REMARKS

I. Explanation of Amendments and Interview Summary

The Applicants acknowledge with thanks the courtesy extended by the Examiner to the applicant's attorney David A. Gass during a personal interview on August 2, 2007, during which time the rejections in the outstanding Office action were discussed. The Applicants proposed the presentation of a meta-analysis of data pertaining to the invention and the Examiner encouraged the Applicants to present such an analysis in the form of a Rule 132 declaration.

The claims were amended to remove reference to "stroke" in order to expedite allowance of a preferred embodiment, and not for reasons related to patentability.

The final clause of claim 61, pertaining to interpretation of a the screening results where the haplotype of interest is *absent*, was amended to clarify that "the absence of the haplotype in the nucleic acid of the individual identifies the individual as not having the elevated susceptibility to MI due to the haplotype. It will be self-evident that any screening test leads to one conclusion if the test is positive and another conclusion if the test is negative. The final clause of the claim as originally presented refers to the relative risk from the tested haplotype and a person of ordinary skill would have interpreted the claim as drawing a conclusion that that the individual has no elevated susceptibility to MI from other, untested factors. That, however, appears to be a conclusion drawn by the Patent Office, resulting in a rejection alleging lack of enablement. The current amendment further clarifies a conclusion that would be reached for a subject that is found *not* to carry the haplotype. Any further fine-tuning of this clause should be amenable to resolution by telephonic interview because the Applicants and the Patent Office appear to intend the same scope and meaning for this element of the claim.

II. Remarks Relating to the nature of the invention

The Applicants continue to dispute the Patent Office's characterization of the claimed invention. The invention relates to a method for assessing susceptibility to myocardial infarction that involves analysis of nucleic acid sequence in a person's FLAP gene. The elected claims are not drawn to polynucleotides, to polymorphisms, or to "differences," even though the Patent Office's patentability search may involve looking for such features in the prior art. Rather, the claims are drawn to methods that involve analyzing a human individual's DNA at a particular locus. The results of the analysis determine whether or not the individual is scored as having elevated risk for myocardial infarction. There is common utility for all variations of this method that are described in the application.

III. The Rejection Under 35 U.S.C. § 112, First Paragraph, Alleging Lack of Enablement Should be Withdrawn

In paragraph 6 the of the Office action the Patent Office rejected claims 61 and 63-66, alleging lack of enabling disclosure. The Applicants traverse this rejection.

The Applicants repeat by reference arguments made in their previous submissions.

The rejection was based in part on alleged overbreadth insofar as the claims encompassed assessing susceptibility to both MI *and stroke*. Reference to stroke has been deleted by amendment, rendering moot this basis for rejection.

Most of the remaining discussion of the issue of enabling disclosure in the Office action focuses on whether or not the association between FLAP haplotypes and MI taught in the application is reproducible. Accompanying this amendment is a sworn declaration summarizing a meta-analysis conducted by deCODE genetics, the assignee of this application. The meta-analysis shows that the correlation between FLAP haplotypes and MI is indeed reproducible.

Importantly, the meta-analysis includes data from the Zee study cited by the Examiner as well as other published studies with available data analyzing the correlation between the FLAP haplotype and MI. The meta-analysis is an aggregation of data from smaller studies and shows that the correlation between the FLAP haplotype and increased risk for MI is real and statistically significant. The statistical power of the meta-analysis is much greater due to the large sample size and is much more probative that any individual smaller study that did not necessarily detect the correlation due to small sample size.

The Patent Office analyzes various articles or studies that pertain in some fashion to gene-disease correlation, but that are *unrelated* to the subject matter of the claims (e.g., Mayer et al., SNP's in the CADPKL gene and neurological disorders). These studies have no probative value with respect to FLAP-MI correlation, especially in view of the abundant data now available pertaining to FLAP-MI. The Applicants pointed to numerous defects in articles cited by the Patent Office or their failure to support the proposition for which they were cited, yet the PTO has continued to rely on the articles without addressing the defects.

The Patent Office alleges (section titled "Guidance in the Specification") that the specification provides no evidence that the invention can be practiced "as broadly claimed" with respect to individual markers. This aspect of the Office action has no relevance to the elected claims, which pertain to a particular four-polymorphism haplotype. The haplotype of the claim is shown to have a statistically significant correlation with increased risk of MI in the application and in the meta-analysis referred to above.

The Patent Office raised concerns about the proper wording of the claims as they pertain to individuals that do not have the tested-for HapA FLAP haplotype. These concerns are addressed above, and do not give rise to any questions of enabling disclosure.

The Patent Office's final concern appears to relate to ethnic or "inter-ethnic" variability conferring different risks. Even if true, such variability does not give rise to questions of statutory enablement for the present claims. Statutory enablement involves whether an application describes an invention in a manner that allows those of ordinary skill to practice the invention. The present application teaches a person of ordinary skill how to perform the haplotype screen without regard to ethnicity, and teaches the conclusion that can reasonably be drawn from it based on population genetics. As with other correlation tests, the results provide helpful information for medical treatment or lifestyle management, and are indicative of risk at the population level. The present invention is appropriately claimed insofar as an individual is assessed for one type of data (FLAP haplotype) and a conclusion about susceptibility (supported by statistically validated data) is drawn based on the FLAP haplotype assessment only. The conclusion does not require ethnicity data.

Human variability is the rule, not the exception, for all aspects of medicine, including diagnostic tests based on biochemistry; safety of drugs; efficacy of drugs; susceptibility to diseases; life expectancy, and so on. While it may be possible or desirable to refine any medical test or treatment or other medical procedure to an ethnicity or sub-ethnicity, that is not the current state of medicine and is not part of the statutory requirement of enablement for a claim that does not require a conclusion based on ethnicity. The

Applicants previous amendment cited many examples of diagnostic tests that are considered medically useful, even though their predictive value with respect to any particular person is not considered a certainty. The data in the application and the larger meta-analysis show that the test is valid and useful and provides another tool for assessing risk for MI.

The Patent Office alleges that "as a general rule of thumb" the field looks for a relative risk of three or more before accepting a paper for publication. As explained in greater detail in the declaration filed herewith, RR $\geq$ 3 is clearly NOT the standard in the field for accepting publications. (The inventor's paper was published in prestigious *Nature* Genetics without RR $\geq$ 3.) Nor is RR $\geq$ 3 descriptive of the risk that would reasonably be attributed to multi-factorial diseases. Nor is RR $\geq$ 3 a relevant indicator for statutory enablement. A conclusion of enablement is appropriate in view of the fact that the incremental risk for MI associated with FLAP HapA, though not nearly as high as 3.0, has been shown through large studies and meta-analysis of multiple studies to be statistically significant, and not an artifact of a small study.

The Patent Office observed that the Helgadottir declaration inadvertently used the term "co-inventor" when "inventor" should have been used. The Applicants apologize for any confusion caused by this typographical error. In addition, the filed Helgadottir declaration omitted Exhibit G, which is references Falk and Rubinstein and Terwilliger and Ott. These references are submitted herewith as Appendix B.

CONCLUSION

In view of the foregoing amendment and remarks, Applicants believe pending claims 61-66 are in condition for allowance and early notice thereof is solicited.

Dated: October 16, 2007

Respectfully submitted,

By
Sharon M. Sintich
Registration No.: 48,484
MARSHALL, GERSTEIN & BORUN LLP
233 S. Wacker Drive, Suite 6300
Sears Tower
Chicago, Illinois 60606-6357
(312) 474-6300
Attorney for Applicant

nature genetics

The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke

Anna Helgadottir[1], Andrei Manolescu[1], Gudmar Thorleifsson[1], Solveig Gretarsdottir[1], Helga Jonsdottir[1], Unnur Thorsteinsdottir[1], Nilesh J Samani[2], Gudmundur Gudmundsson[1], Struan F A Grant[1], Gudmundur Thorgeirsson[3], Sigurlaug Sveinbjornsdottir[3], Einar M Valdimarsson[3], Stefan E Matthiasson[3], Halldor Johannsson[3], Olof Gudmundsdottir[1], Mark E Gurney[1], Jesus Sainz[1], Margret Thorhallsdottir[1], Margret Andresdottir[1], Michael L Frigge[1], Eric J Topol[4], Augustine Kong[1], Vilmundur Gudnason[5], Hakon Hakonarson[1], Jeffrey R Gulcher[1] & Kari Stefansson[1]

We mapped a gene predisposing to myocardial infarction to a locus on chromosome 13q12–13. A four-marker single-nucleotide polymorphism (SNP) haplotype in this locus spanning the gene *ALOX5AP* encoding 5-lipoxygenase activating protein (FLAP) is associated with a two times greater risk of myocardial infarction in Iceland. This haplotype also confers almost two times greater risk of stroke. Another *ALOX5AP* haplotype is associated with myocardial infarction in individuals from the UK. Stimulated neutrophils from individuals with myocardial infarction produce more leukotriene B4, a key product in the 5-lipoxygenase pathway, than do neutrophils from controls, and this difference is largely attributed to cells from males who carry the at-risk haplotype. We conclude that variants of ALOX5AP are involved in the pathogenesis of both myocardial infarction and stroke by increasing leukotriene production and inflammation in the arterial wall.

Cardiovascular diseases (CVD) are the leading causes of death and disability in the developed world[1], with an increasing prevalence due to the aging of the population and the obesity epidemic. More than 1 million deaths in the US alone were caused by myocardial infarction and stroke in 2003 (ref. 2). Some of the processes underlying myocardial infarction are now understood: it is generally attributed to atherosclerosis with arterial wall inflammation that ultimately leads to plaque rupture, fissure or erosion[3,4]. This process is known to involve diapedesis of monocytes across the endothelial barrier; activation of neutrophils, macrophage cells and platelets; and release of a variety of cytokines and chemokines[5,6], but the genetic basis of the process has not yet been deciphered.

Two different approaches have been used to search for genes associated with myocardial infarction. SNPs in candidate genes have been tested for association and have, in general, not been replicated or confer only a modest risk of myocardial infarction. Case-control association studies have identified several proinflammatory genes with variants that are associated with either an increased risk of myocardial infarction or a protective effect[7–9]. Four genome-wide scans in families with myocardial infarction have yielded several loci with formidable linkage peaks, but the gene(s) underlying these loci have not yet been identified[10–14]. In addition, one large pedigree study identified a deletion mutation of a transcription factor gene, *MEF2A*, with autosomal dominant transmission[14]. This is an interesting cause of myocardial infarction, but the prevalence of this or other mutations in *MEF2A* outside this family remains to be determined.

Here we report a genome-wide scan of 296 multiplex Icelandic families including 713 individuals with myocardial infarction. Through suggestive linkage to a locus on chromosome 13q12–13, we identified the gene (*ALOX5AP*) encoding FLAP and found that a four-SNP haplotype in the gene confers a nearly two times greater risk of myocardial infarction and stroke. FLAP is a regulator[15] of a crucial pathway in the genesis of leukotriene inflammatory mediators, which are implicated in atherosclerosis both in a mouse model[16] and in human studies[17,18]. Males had the strongest association to the at-risk haplotype, and male carriers of the at-risk haplotype also had significantly greater production of leukotriene-B4 (LTB4), supporting the idea that proinflammatory activity has a role in the pathogenesis of myocardial infarction. We confirmed the association of *ALOX5AP* with myocardial infarction in an independent cohort of British individuals with another haplotype. These results indicate that *ALOX5AP* is the first specific gene isolated that confers substantial population-attributable risk (PAR) of the complex traits of both myocardial infarction and stroke.

[1]deCODE genetics, Sturlugata 8, Reykjavik, Iceland. [2]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Leicester, UK. [3]National University Hospital, Reykjavik, Iceland. [4]Cleveland Clinic Foundation, Cleveland, Ohio, USA. [5]Icelandic Heart Association, Reykjavik, Iceland. Correspondence should be addressed to K.S. (kstefans@decode.is).

Published online 8 February 2004; doi:10.1038/ng1311

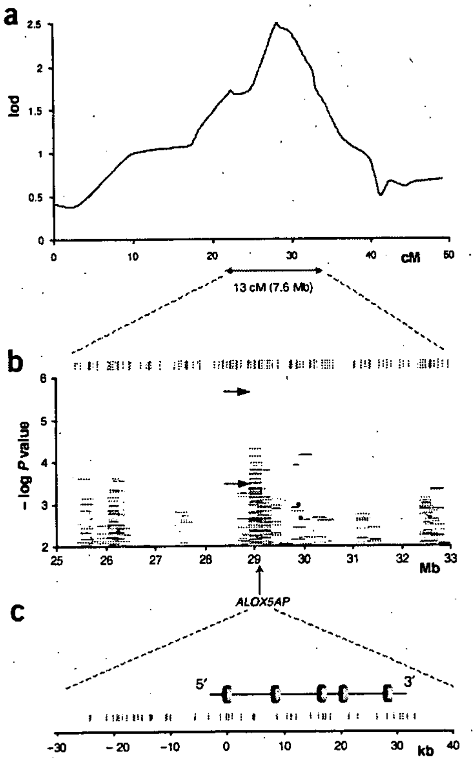


Figure 1 Schematic view of the chromosome 13 linkage region showing *ALOX5AP*. **(a)** The linkage scan for females with myocardial infarction and the one-lod drop region that includes *ALOX5AP*. **(b)** Microsatellite association for all individuals with myocardial infarction: single-marker association (black dots) and two-, three-, four- and five-marker haplotype association (black, blue, green and red horizontal lines, respectively). The blue and red arrows indicate the location of the most significant haplotype association across *ALOX5AP* in males and females, respectively. **(c)** *ALOX5AP* gene structure, with exons shown as colored cylinders, and the locations of all SNPs typed in the region. The green vertical lines indicate the position of the microsatellites **(b)** and SNPs **(c)** used in the analysis.

RESULTS

Linkage analysis

We carried out a genome-wide scan in search of myocardial infarction susceptibility genes using a framework set of 1,068 microsatellite markers. The initial linkage analysis included 713 individuals with myocardial infarction who fulfilled the World Health Organization (WHO) MONICA research criteria[19] and were clustered in 296 extended families. We repeated the linkage analysis for individuals with early onset, for males and for females separately. A description of the number of affected individuals and families in each analysis is provided in **Supplementary Table 1** online, and the corresponding allele-sharing lod scores are given in **Supplementary Figure 1** online. None of these analyses yielded a locus of genome-wide significance. The most promising lod score (2.86) was observed on chromosome 13q12–13 for linkage with females with myocardial infarction at the peak marker *D13S289* (**Supplementary Fig. 1** online). This locus also had the most promising lod score (2.03) for individuals with early-onset myocardial infarction. After we increased the information on identity-by-descent sharing to over 90% by typing an additional 14 microsatellite markers in a 30-cM region around *D13S289*, the lod score for the association in females dropped to 2.48 ($P = 0.00036$), and the lod score remained highest at *D13S289* (**Fig. 1a**). In an independent linkage study of males with ischemic stroke or transient ischemic attack (TIA), we observed linkage to the same locus with a lod score of 1.51 at the same peak marker (**Supplementary Fig. 2** online), further suggesting that a cardiovascular susceptibility factor might reside at this locus.

Microsatellite association study

The 7.6-Mb region that corresponds to a drop of 1 in lod score in the female–myocardial infarction linkage analysis contains 40 known genes (**Supplementary Table 2** online). To determine which gene in this region was most likely to contribute to myocardial infarction, we typed 120 microsatellite markers in the region and carried out a case-control association study using 802 unrelated (separated by at least three meioses) individuals with myocardial infarction and 837 population-based controls. We also repeated the association study for each of the three phenotypes that were used in the linkage study: individuals with early onset, males and females with myocardial infarction. In addition to testing each marker individually, we also tested haplotypes based on these markers for association. To limit the number of haplotypes tested, we considered only haplotypes spanning less than 300 kb that were over-represented among the affected individuals.

The haplotype with the strongest association to myocardial infarction ($P = 0.00004$) covered a region that contains two known genes: *ALOX5AP* (**Fig. 1b**) and a gene with an unknown function called highly charged protein (*D13S106E*). The haplotype association in this region for females with myocardial infarction was less significant ($P = 0.0004$) than for all individuals with myocardial infarction, and the most significant haplotype association was observed for males with myocardial infarction ($P = 0.000002$). The haplotype associated with males with myocardial infarction was the only haplotype that retained significant association after adjusting for all haplotypes tested.

FLAP, together with 5-lipoxygenase (5-LO), is a regulator of the leukotriene biosynthetic pathway that has recently been implicated in the pathogenesis of atherosclerosis[16–18]. Therefore, *ALOX5AP* was a good candidate for the gene underlying the association with myocardial infarction.

Screening for SNPs in *ALOX5AP* and LD mapping

To determine whether variations in *ALOX5AP* significantly associate with myocardial infarction and to search for causal variations, we sequenced *ALOX5AP* in 93 affected individuals and 93 controls. The sequenced region covers 60 kb containing *ALOX5AP*, including the five known exons and introns, the 26-kb region 5′ to the first exon and the 7-kb region 3′ to the fifth exon. We identified 144 SNPs, of which we excluded 96 from further analysis owing to either a low minor allele frequency or complete correlation (redundancy) with other SNPs. **Figure 1c** shows the distribution of the 48 SNPs chosen for genotyping, relative to exons, introns and the 5′ and 3′ flanking regions of *ALOX5AP*. We identified only one SNP in a coding sequence (exon 2), which did not lead to an amino acid substitution. The locations of the 48 SNPs in the National Center for Biotechnology Information human genome assembly build 34 are listed in **Supplementary Table 3** online. In addition to the SNPs, we typed a polymorphism consisting of a monopolymer A repeat in the *ALOX5AP* promoter region[20].

The linkage disequilibrium (LD) block structure defined by the 48 genotyped SNPs is shown in **Figure 2**. Strong LD was detected across the *ALOX5AP* region, although at least one historical recombination seems to have occurred, dividing the region into two strongly correlated LD blocks.

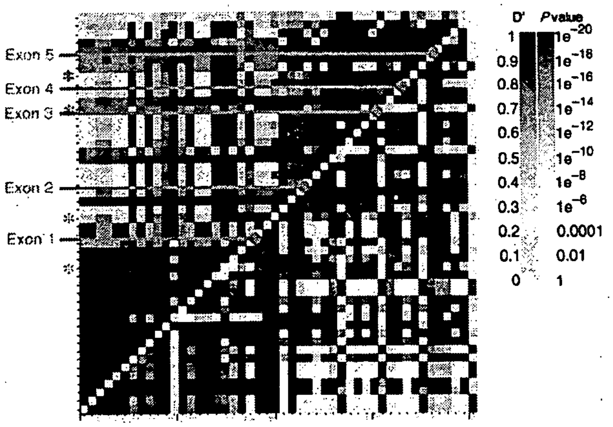


Figure 2 Pairwise LD between SNPs in a 60-kb region encompassing *ALOX5AP*. The markers are plotted equidistantly. Two measures of LD are shown: *D′* in the upper left triangle and *P* values in the lower right triangle. Colored lines indicate the positions of the exons of *ALOX5AP*, and the green stars indicate the location of the markers of the at-risk haplotype HapA. Scales for both measures of the LD strength are provided on the right.

Haplotype association with myocardial infarction

In a case-control association study, we genotyped the 48 selected SNPs and the monopolymer A repeat marker in a set of 779 unrelated individuals with myocardial infarction and 624 population-based controls. We tested each of the 49 markers individually for association with the disease. Three SNPs, one located 3 kb upstream of the first exon and the other two 1 kb and 3 kb downstream of the first exon, showed nominally significant association to myocardial infarction (**Supplementary Table 4** online). After adjusting for the number of markers tested, however, these results were not significant. We then searched for haplotypes associated with the disease using the same cohorts. We limited the search to haplotype combinations constructed from two, three or four SNPs and tested only haplotypes that were over-represented in the individuals with myocardial infarction. The resulting *P* values were adjusted for all the haplotypes we tested by randomizing the affected individuals and controls.

Several haplotypes were significantly associated with the disease at an adjusted significance level of $P < 0.05$ (**Supplementary Table 5** online). We observed the most significant association with a four-SNP haplotype spanning 33 kb, including the first four exons of *ALOX5AP* (**Fig. 1c**), with a nominal *P* value of 0.0000023 and an adjusted *P* value of 0.005. This haplotype, called HapA, has a haplotype frequency of 15.8% (carrier frequency 29.1%) in affected individuals versus 9.5% (carrier frequency 18.1%) in controls (**Table 1**). The relative risk conferred by HapA compared with other haplotypes constructed from the same SNPs, assuming a multiplicative model, was 1.8 and the corresponding PAR was 13.5%. HapA was present at a higher frequency in males (carrier frequency 30.9%) than in females with myocardial infarction (carrier frequency 25.7%; **Table 1**). All other haplotypes that were significantly associated with an adjusted *P* value less than 0.05 were highly correlated with HapA and should be considered variants of that haplotype (**Supplementary Table 5** online).

Association of HapA with stroke and PAOD

Because of the high degree of comorbidity among myocardial infarction, stroke and peripheral arterial occlusive disease (PAOD), with most of these cases occurring on the basis of an atherosclerotic disease, we wanted to determine whether HapA was also associated with stroke or PAOD. We typed the SNPs defining HapA for these cohorts. We removed first- and second-degree relatives and all known cases of myocardial infarction and tested for association in 702 individuals with stroke and 577 individuals with PAOD (**Table 1**). We observed a significant association of HapA with stroke, with a relative risk of 1.67 ($P = 0.000095$). In addition, we determined whether HapA was primarily associated with a particular subphenotype of stroke and found that both ischemic and hemorrhagic stroke were significantly associated with HapA (**Supplementary Table 6** online). Finally, although HapA was more frequent in the PAOD cohort than in the population controls (**Table 1**), this was not significant. Similar to the stronger association of HapA with males with myocardial infarction than with females with myocardial infarction, HapA also showed stronger association with males than with females with stroke and PAOD (**Table 1**).

Haplotype association in a British cohort

In an independent study, we determined whether variants in *ALOX5AP* also affected the risk of myocardial infarction in a population outside Iceland. We typed SNPs defining HapA in a cohort of 753 individuals from the UK who had sporadic myocardial infarction and in 730 British population controls. The affected individuals and controls were from three separate study cohorts recruited in Leicester and Sheffield. We found a slightly higher frequency of HapA in affected individuals versus controls (16.8% versus 15.1%, respectively), but the results were not statistically significant. As in the Icelandic population, HapA was more common in males with myocardial infarction (carrier frequency 31.7%) than in females with myocardial infarction (carrier frequency 28.0%). When we typed an additional nine SNPs, distributed across *ALOX5AP*, in the British cohort and searched for other haplotypes that might be associated with myocardial infarction, two SNPs showed association to myocardial infarction with a nominally significant *P* value (data not shown). Moreover, three- and four-SNP haplotype combinations were associated with higher risk of myocardial infarction in the British cohort, and we observed the most signifi-

Table 1 Association of HapA with myocardial infarction, stroke and PAOD

Phenotype (*n*)	Frequency	RR	PAR	*P* value	*P* value[a]
Myocardial infarction (779)	0.158	1.80	0.135	0.0000023	0.005
Males (486)	0.169	1.95	0.158	0.00000091	ND
Females (293)	0.138	1.53	0.094	0.0098	ND
Early onset (358)	0.139	1.53	0.094	0.0058	ND
Stroke (702)[b]	0.149	1.67	0.116	0.000095	ND
Males (373)	0.156	1.76	0.131	0.00018	ND
Females (329)	0.141	1.55	0.098	0.0074	ND
PAOD (577)[b]	0.122	1.31	0.056	0.061	ND
Males (356)	0.126	1.36	0.065	0.057	ND
Females (221)	0.114	1.22	0.041	0.31	ND

[a] *P* value adjusted for the number of haplotypes tested. [b] Excluding known cases of myocardial infarction.

Shown is HapA of *ALOX5AP* and the corresponding number of affected individuals (*n*), the haplotype frequency in affected individuals, the relative risk (RR), PAR and *P* values. HapA is defined by the SNPs SG13S25, SG13S114, SG13S89 and SG13S32 (**Supplementary Table 5** online). The same controls ($n = 624$) were used for the association analysis in myocardial infarction, stroke and PAOD as well as for the analysis of males, females and individuals with early onset. The frequency of HapA in the control cohort is 0.095. ND, not done.

Table 2 Association of HapB with myocardial infarction in British individuals

Phenotype (*n*)	Frequency	RR	PAR	*P* value	*P* value[a]
Myocardial infarction (753)	0.075	1.95	0.072	0.00037	0.046
Males (549)	0.075	1.97	0.072	0.00093	ND
Females (204)	0.073	1.90	0.068	0.021	ND

[a] *P* value adjusted for the number of haplotypes tested using 1,000 randomization tests.

Shown are the results for HapB that shows the strongest association in the British myocardial infarction cohort. HapB is defined by the SNPs SG13S377, SG13S114, SG13S41 and SG13S35, which have the alleles A, A, A and G, respectively. In all three phenotypes shown, the same set of 730 British controls was used and the frequency of HapB in the control cohort is 0.040. Number of affected individuals (*n*), haplotype frequency in affected individuals, relative risk (RR) and PAR are indicated. ND, not done.

cant association for a four-SNP haplotype with a nominal *P* value of 0.00037 (**Table 2**). We call this haplotype HapB. The haplotype frequency of HapB was 7.5% in the individuals with myocardial infarction (carrier frequency 14.4%) compared with 4.0% (carrier frequency 7.8%) in controls, conferring a relative risk of 1.95 (**Table 2**). This association of HapB remained significant after adjusting for all haplotypes tested, using 1,000 randomization steps, with an adjusted $P = 0.046$. No other SNP haplotype had an adjusted *P* value <0.05. The two at-risk haplotypes, HapA and HapB, are mutually exclusive; there are no instances in which the same chromosome carries both haplotypes.

More LTB4 in individuals with myocardial infarction

To determine whether individuals with a past history of myocardial infarction had greater activity of the 5-LO pathway than controls, we measured production of LTB4 (a key product of the 5-LO pathway) in blood neutrophils isolated from Icelandic individuals with myocardial infarction and controls before and after stimulation with the calcium ionophore ionomycin. We detected no difference in LTB4 production in resting neutrophils from individuals with myocardial infarction versus controls. In contrast, LTB4 generation by neutrophils stimulated with ionomycin was substantially greater in individuals with myocardial infarction than in controls after 15 and 30 min, respectively (**Fig. 3a**). Moreover, the observed difference in release of LTB4 was largely accounted for by male carriers of HapA (**Fig. 3b**), whose cells produced significantly more LTB4 than cells from controls ($P = 0.0042$; **Supplementary Table 7** online). There was also a heightened LTB4 response in males who did not carry HapA, but this difference was of borderline significance (**Supplementary Table 7** online). This could be explained by additional variants in *ALOX5AP* that have not been uncovered, or in other genes belonging to the 5-LO pathway, that may account for upregulation of the LTB4 response in some individuals without the *ALOX5AP* at-risk haplotype. We did not detect differences in LTB4 response in females (**Supplementary Table 7** online), but because of the small sample size, this result is not conclusive. The elevated levels of LTB4 production in stimulated neutrophils from male carriers of the at-risk haplotype suggest that the disease-associated variants of *ALOX5AP* heighten the response of FLAP to factors that stimulate inflammatory cells.

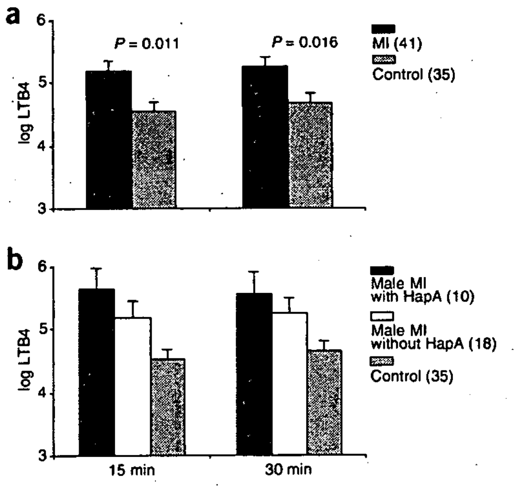


Figure 3 LTB4 production of ionomycin-stimulated neutrophils from individuals with myocardial infarction ($n = 41$) and controls ($n = 35$). The log-transformed (mean ± s.d.) values measured at 15 and 30 min in stimulated cells are shown. (**a**) LTB4 production in individuals with myocardial infarction (MI) and controls. The difference in the mean values between affected individuals and controls was tested using a two-sample *t*-test of the log-transformed values. (**b**) LTB4 production in males with myocardial infarction carrying HapA (red bars) and not carrying HapA (white bars). Mean values of controls (blue bars) are included for comparison. Males with HapA produced the highest amounts of LTB4 ($P < 0.005$ compared with controls). Data for females are shown in **Supplementary Table 7** online.

DISCUSSION

Our results show that variants of *ALOX5AP* encoding FLAP are associated with greater risk of myocardial infarction and stroke. In our Icelandic cohort, a haplotype that spans *ALOX5AP* is carried by 29.1% of all individuals with myocardial infarction and almost doubles the risk of myocardial infarction. We then replicated these findings in an independent cohort of individuals with stroke. Furthermore, stimulated neutrophils from individuals with myocardial infarction had greater production of LTB4, one of the key products of the 5-LO pathway. When we examined this in the context of the at-risk haplotype, however, the gain of function was largely attributed to male carriers of the at-risk haplotype, who also had the strongest association with the *ALOX5AP* haplotype. Another haplotype spanning *ALOX5AP* was associated with myocardial infarction in a British cohort. Although the pathogenic variants responsible for the effects associated with the disease haplotypes are unknown, the greater production of LTB4 observed in ionomycin-stimulated neutrophils from male carriers of the at-risk haplotype suggests that the disease-associated variants increase the response of FLAP to factors that stimulate inflammatory cells.

We observed suggestive linkage to chromosome 13q12–13 with several different phenotypic groups, including females with myocardial infarction, individuals of both sexes with early-onset myocardial infarction and males with ischemic stroke or TIA. But we observed the strongest haplotype association for males with myocardial infarction or stroke. Therefore, the linkage signal in females with myocardial infarction and in individuals with early-onset myocardial infarction is not explained by the at-risk haplotype that we identified, and we expect that there may be other unidentified variants or haplotypes in *ALOX5AP*, or in other genes in the linkage region, that may confer risk of these cardiovascular phenotypes. These variants are probably rarer than HapA with relatively high penetrance, higher in women than in men.

FLAP has an important role in the initial steps of leukotriene biosynthesis[15], which is largely confined to leukocytes and can be triggered by a variety of stimuli. In this biosynthetic pathway, unesterified arachidonic acid is converted to LTA4 by the action of 5-LO and its activating protein FLAP. The unstable epoxide LTA4 is further metabolized to LTB4 or LTC4 by LTA4 hydrolase and LTC4 synthase, respectively. In addition, LTA4 can be exported to neighboring cells that are devoid of 5-LO activity and become subject to transcellular leukotriene biosynthesis[21–23]. The leukotrienes have a variety of proinflammatory effects[24,25]. LTB4 activates leukocytes, leading to chemotaxis and increased adhesion of leukocytes to vascular endothelium, release of lysosomal enzymes such as myeloperoxidase and production of superoxide anions[25]. The cysteinyl-containing leukotrienes (LTC4 and its metabolites LTD4 and LTE4) increase vascular permeability in postcapillary venules and are potent vasoconstrictors of coronary arteries[26–28].

The importance of the 5-LO pathway is well established in asthma, and drugs inhibiting this pathway have been developed for treating asthma. The role of the 5-LO pathway in the pathogenesis of atherosclerosis has recently received attention. A study of postmortem pathologic specimens showed an increase in the expression of members of the 5-LO pathway, including 5-LO and FLAP, in atherosclerotic lesions at various stages of development in the aorta, coronary arteries and carotid arteries[18]. Furthermore, 5-LO was localized to macrophages, dendritic cells, foam cells, mast cells and neutrophilic granulocytes, and the number of cells expressing 5-LO was markedly greater in advanced lesions[18]. The leukocytes positive for 5-LO accumulated at distinct sites that are most prone to rupture[29], such as the shoulder regions below the fibrous cap of the atherosclerotic lesion[18]. A 5-LO promoter variant is associated with abnormal carotid artery intima-media thickness and heightened inflammatory biomarkers[30]. In addition, antagonists of LTB4 block the development of atherosclerosis in apo-E-deficient and LDRL-deficient mice[31], and a congenic mouse strain with a heterozygous deficiency of 5-LO shows resistance to atherosclerosis[16], further supporting the idea that greater activity of the 5-LO pathway has a role in predisposition to atherosclerosis.

Our data also show that the at-risk haplotype of *ALOX5AP* has higher frequency in all subgroups of stroke, including ischemic stroke, TIA and hemorrhagic stroke. HapA confers significantly higher risk of myocardial infarction and stroke than it does of PAOD. This could be explained by differences in the pathogenesis of these diseases. Unlike individuals with PAOD, who have ischemic legs because of atherosclerotic lesions that are responsible for gradually diminishing blood flow to the legs, individuals with myocardial infarction and stroke have suffered acute events, with disruption of the vessel wall suddenly decreasing blood flow to regions of the heart and the brain.

We did not find association between HapA and myocardial infarction in a British cohort, but we did find significant association between myocardial infarction and a different *ALOX5AP* variant. The existence of different haplotypes of the gene conferring risk to myocardial infarction in different populations is not unexpected. It is not unreasonable to assume that a common disease like myocardial infarction is associated with many different mutations or sequence variations and that the frequencies of these disease-associated variants may differ between populations. It would also not be unexpected for the same mutation to arise on different haplotypic backgrounds.

Our work suggests that *ALOX5AP* has an important role in the pathogenesis of myocardial infarction and stroke in humans. Our study, together with others, may provide the necessary background to launch therapeutic trials to determine whether pharmacological inhibition of FLAP will prevent the development of myocardial infarction and stroke.

METHODS

Study population. We recruited the individuals in the study from a registry of over 8,000 individuals, which includes all individuals who had myocardial infarctions before the age of 75 in Iceland from 1981 to 2000. This registry is a part of the WHO MONICA Project[19]. Diagnoses of all individuals in the registry follow strict diagnostic rules based on signs, symptoms, electrocardiograms, cardiac enzymes and necropsy findings.

We used genotypes from 713 individuals with myocardial infarction and 1,741 of their first-degree relatives in the linkage analysis. For the microsatellite association study of the locus associated with myocardial infarction, we used 802 unrelated (no first- or second-degree relatives) individuals with myocardial infarction (233 females, 624 males and 302 with early onset) and 837 population-based controls. The females studied were post-menopausal. Over 90% of the individuals were taking aspirin or other nonsteroidal anti-inflammatory drugs. For the SNP association study in and around *ALOX5AP*, we genotyped 779 unrelated individuals with myocardial infarction (293 females, 486 males and 358 with early onset). The control group for the SNP association study was population-based and comprised of 624 unrelated males and females 20–90 years of age whose medical history was unknown. The stroke and PAOD cohorts used in this study have previously been described[32–34]. For the stroke linkage analysis, we used genotypes from 342 males with ischemic stroke or TIA that were linked to at least one other male within and including six meioses in 164 families. For the association studies, we analyzed 702 individuals with all forms of stroke (329 females and 373 males) and 577 individuals with PAOD (221 females and 356 males). Individuals with stroke or PAOD who also had myocardial infarction were excluded. Controls used for the stroke and PAOD association studies were the same as used in the myocardial infarction SNP association study.

The study was approved by the Data Protection Commission of Iceland and the National Bioethics Committee of Iceland. We obtained informed consent from all study participants. Personal identifiers associated with medical information and blood samples were encrypted with a third-party encryption system as previously described[35].

Statistical analysis. We carried out a genome-wide scan as previously described[33], using a set of 1,068 microsatellite markers. We used multipoint, affected-only allele-sharing methods[36] to assess the evidence for linkage. All results were obtained using the program Allegro[37] and the deCODE genetic map[38]. We used the S_{pairs} scoring function[39,40] and the exponential allele-sharing model[36] to generate the relevant 1-degree-of-freedom statistics. When combining the family scores to obtain an overall score, we used a weighting scheme that is halfway on a log scale between weighting each affected pair equally and weighting each family equally. In the analysis, all genotyped individuals who were not affected were treated as 'unknown'. Because of concern with small-sample behavior, we usually computed corresponding *P* values in two different ways for comparison and report the less significant one. The first *P* value was computed based on large sample theory, $Z_{lr} = \sqrt{(2 \log_e(10) \text{lod})}$, and is distributed approximately as a standard normal distribution under the null hypothesis of no linkage[36]. A second *P* value was computed by comparing the observed lod score with its complete data sampling distribution under the null hypothesis[37]. When a data set consisted of more than a handful of families, these two *P* values tended to be very similar. The information measure we used, which is implemented in Allegro, is closely related to a classical measure of information and has a property that is between 0 (if the marker genotypes are completely uninformative) and 1 (if the genotypes determine the exact amount of allele sharing by descent among the affected relatives)[41,42].

For single-marker association studies, we used Fisher's exact test to calculate two-sided *P* values for each allele. All *P* values are unadjusted for multiple comparisons unless specifically indicated. We present allelic rather than carrier frequencies for microsatellites, SNPs and haplotypes. To minimize any bias due to the relatedness of the individuals who were recruited as families for the linkage analysis, we eliminated first- and second-degree relatives. For the haplotype analysis we used the program NEMO[32], which handles missing genotypes and uncertainty with phase through a likelihood procedure, using the expectation-maximization algorithm as a computational tool to estimate haplotype frequencies. Under the null hypothesis, the affected individuals and controls were assumed to have identical haplotype frequencies. Under the alternative hypotheses, the candidate at-risk haplotype was allowed to have a higher frequency in the affected individuals than in controls, and the ratios of frequencies of all other haplotypes were assumed to be the same in both groups. Likelihoods were maximized separately under both hypotheses, and a corresponding 1-degree-of-freedom likelihood ratio statistic was used to evaluate statistical significance[32]. Although we only searched for haplotypes that increased the risk, all reported *P* values are two-sided unless otherwise stated. To assess the significance of the haplotype association corrected for multiple testing, we carried out a randomization test using the same genotype data. We randomized the cohorts of affected individuals and controls and repeated the analysis. This procedure was repeated up to 1,000 times, and the *P* value we present is the fraction of replications that produced a *P* value for a haplotype tested that was lower than or equal to the *P* value we observed using the original affected individual and control cohorts.

For both single-marker and haplotype analysis, we calculated relative risk (RR) and PAR assuming a multiplicative model[43,44] in which the risk of the two alleles of haplotypes a person carries multiply. We calculated LD between pairs of SNPs using the standard definition of D' (ref. 45) and R^2 (ref. 46). Using NEMO, we estimated frequencies of the two marker allele combinations by maximum likelihood and evaluated deviation from linkage equilibrium by a likelihood ratio test. When plotting all SNP combinations to elucidate the LD structure in a particular region, we plotted D' in the upper left corner and the *P* value in the lower right corner. In the LD plots we present, the markers are plotted equidistantly rather than according to their physical positions.

Identification of DNA polymorphisms. We identified new polymorphic repeats (dinucleotide or trinucleotide repeats) with the Sputnik program. We subtracted the lower allele of the CEPH sample 1347-02 (CEPH genomics repository) from the alleles of the microsatellites and used it as a reference. We detected SNPs in the gene by PCR sequencing exonic and intronic regions from affected individuals and controls. We also detected public polymorphisms by BLAST search of the National Center for Biotechnology Information SNP database. We genotyped SNPs using a method for detecting SNPs with fluorescent polarization template-directed dye-terminator incorporation[47] and TaqMan assays (Applied Biosystems).

Isolation and activation of peripheral blood neutrophils. We drew 50 ml of blood from each of 41 individuals with myocardial infarction and 35 age- and sex-matched controls into vacutainers containing EDTA. All blood was drawn at the same time in the early morning after 12 h of fasting. We isolated neutrophils using Ficoll-Paque PLUS (Amersham Biosciences).

We collected the red cell pellets from the Ficoll gradient and then lysed red blood cells in 0.165 M ammonium chloride for 10 min on ice. After washing them with phosphate-buffered saline, we counted neutrophils and plated them at 2×10^6 cells ml^{-1} in 4-ml cultures of 15% fetal calf serum (GIBCO BRL) in RPMI-1640 medium (GIBCO BRL). We then stimulated cells with maximum effective concentration of ionomycin (1 μM). At 0, 15, 30, 60 min after adding ionomycin, we aspirated 600 μl of culture medium and stored it at –80 °C for the measurement of LTB4 release as described below. We maintained cells at 37 °C in a humidified atmosphere of 5% carbon dioxide–95% air. We treated all samples with indomethasine (1 μM) to block the cyclooxygenase enzyme.

Ionomycin-induced release of LTB4 in neutrophils. We used the LTB4 Immunoassay (R&D systems) to quantify LTB4 concentration in supernatant from cultured ionomycin-stimulated neutrophils. The assay we used is based on the competitive binding technique in which LTB4 present in the testing samples (200 μl) competes with a fixed amount of alkaline phosphatase–labeled LTB4 for sites on a rabbit polyclonal antibody. During the incubation, the polyclonal antibody becomes bound to a goat antibody to rabbit coated onto the microplates. After washing to remove excess conjugate and unbound sample, a substrate solution was added to the wells to determine the bound enzyme activity. We stopped the color development and read the absorbance at 405 nm. The intensity of the color is inversely proportional to the concentration of LTB4 in the sample. Each LTB4 measurement using the LTB4 Immunoassay was done in duplicate.

British study population. We recruited three separate British cohorts as described previously[48,49]. The first two cohorts comprised 549 individuals from among those who were admitted to the coronary care units of the Leicester Royal Infirmary, Leicester (July 1993–April 1994), and the Royal Hallamshire Hospital, Sheffield (November 1995–March 1997), and satisfied the WHO criteria for acute myocardial infarction in terms of symptoms, elevations in cardiac enzymes or electrocardiographic changes[50]. We recruited 532 control individuals in each hospital from adult visitors of individuals with noncardiovascular disease on general medical, surgical, orthopedic and obstetric wards to find subjects representative of the source population from which the affected individuals originated. Individuals who reported a history of coronary heart disease were excluded.

In the third cohort, we recruited 204 individuals retrospectively from the registries of three coronary care units in Leicester. All had suffered a myocardial infarction according to WHO criteria before the age of 50 years. At the time of participation, individuals were at least 3 months from the acute event. The control cohort comprised 198 individuals with no personal or family history of premature coronary heart disease, matched for age, sex and current smoking status with the cases. We recruited control individuals from three primary care practices located in the same geographical area. In all cohorts, individuals were white of Northern European origin. Local research ethics committees approved all the studies, and individuals provided written informed consent for use of samples in genetic studies of coronary artery disease.

URLs. The Sputnik program is available at http://espressosoftware.com/pages/sputnik.jsp. The National Center for Biotechnology Information SNP database is available at http://www.ncbi.nlm.nih.gov/SNP/index.html.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We thank the affected individuals and their families whose contribution made this study possible and the nurses at the Icelandic Heart Association, personnel at the deCODE core facilities, T. Jonsdottir, F. Runarsson, E. Palsdottir, J. Kostic, K. Channer, R. Steeds, R. Singh and P. Braund for their contributions. N.J.S. is supported by the British Heart Foundation.

COMPETING INTERESTS STATEMENT
The authors declare competing financial interests (see the *Nature Genetics* website for details).

Received 7 January; accepted 26 January 2004
Published online at http://www.nature.com/naturegenetics/

1. Bonow, R.O., Smaha, L.A., Smith, S.C. Jr., Mensah, G.A. & Lenfant, C. World Heart Day 2002: the international burden of cardiovascular disease: responding to the emerging global epidemic. *Circulation* **106**, 1602–1605 (2002).
2. *Heart Disease and Stroke Statistics, 2003 Update* (American Heart Association, Dallas, Texas, 2002).
3. Lusis, A.J. Atherosclerosis. *Nature* **407**, 233–241 (2000).
4. Libby, P. Inflammation in atherosclerosis. *Nature* **420**, 868–874 (2002).
5. Stratford, N., Britten, K. & Gallagher, P. Inflammatory infiltrates in human coronary atherosclerosis. *Atherosclerosis* **59**, 271–276 (1986).
6. Poole, J.C. & Florey, H.W. Changes in the endothelium of the aorta and the behaviour of macrophages in experimental atheroma of rabbits. *J. Pathol. Bacteriol.* **75**, 245–251 (1958).
7. Topol, E.J. *et al.* Single nucleotide polymorphisms in multiple novel thrombospondin genes may be associated with familial premature myocardial infarction. *Circulation* **104**, 2641–2644 (2001).
8. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32**, 650–654 (2002).
9. Yamada, Y. *et al.* Prediction of the risk of myocardial infarction from polymorphisms in candidate genes. *N. Engl. J. Med.* **347**, 1916–1923 (2002).
10. Broeckel, U. *et al.* A comprehensive linkage analysis for myocardial infarction and its related risk factors. *Nat. Genet.* **30**, 210–214 (2002).
11. Francke, S. *et al.* A genome-wide scan for coronary heart disease suggests in Indo-Mauritians a susceptibility locus on chromosome 16p13 and replicates linkage with the metabolic syndrome on 3q27. *Hum. Mol. Genet.* **10**, 2751–2765 (2001).
12. Harrap, S.B. *et al.* Genome-wide linkage analysis of the acute coronary syndrome suggests a locus on chromosome 2. *Arterioscler. Thromb. Vasc. Biol.* **22**, 874–878 (2002).
13. Pajukanta, P. *et al.* Two loci on chromosomes 2 and X for premature coronary heart disease identified in early- and late-settlement populations of Finland. *Am. J. Hum. Genet.* **67**, 1481–1493 (2000).
14. Wang, L., Fan, C., Topol, S.E., Topol, E.J. & Wang, Q. Mutation of MEF2A in an inherited disorder with features of coronary artery disease. *Science* **302**, 1578–1581 (2003).

15. Dixon, R.A. *et al.* Requirement of a 5-lipoxygenase-activating protein for leukotriene synthesis. *Nature* **343**, 282–284 (1990).
16. Mehrabian, M. *et al.* Identification of 5-lipoxygenase as a major gene contributing to atherosclerosis susceptibility in mice. *Circ. Res.* **91**, 120–126 (2002).
17. Brezinski, D.A., Nesto, R.W. & Serhan, C.N. Angioplasty triggers intracoronary leukotrienes and lipoxin A4. Impact of aspirin therapy. *Circulation* **86**, 56–63 (1992).
18. Spanbroek, R. *et al.* Expanding expression of the 5-lipoxygenase pathway within the arterial wall during human atherogenesis. *Proc. Natl. Acad. Sci. USA* **100**, 1238–1243 (2003).
19. The World Health Organization MONICA Project (monitoring trends and determinants in cardiovascular disease): a major international collaboration. WHO MONICA Project Principal Investigators. *J. Clin. Epidemiol.* **41**, 105–14 (1988).
20. Koshino, T. *et al.* Novel polymorphism of the 5-lipoxygenase activating protein (FLAP) promoter gene associated with asthma. *Mol. Cell. Biol. Res. Commun.* **2**, 32–35 (1999).
21. Sala, A., Bolla, M., Zarini, S., Muller-Peddinghaus, R. & Folco, G. Release of leukotriene A4 versus leukotriene B4 from human polymorphonuclear leukocytes. *J. Biol. Chem.* **271**, 17944–17948 (1996).
22. Dahinden, C.A., Clancy, R.M., Gross, M., Chiller, J.M. & Hugli, T.E. Leukotriene C4 production by murine mast cells: evidence of a role for extracellular leukotriene A4. *Proc. Natl. Acad. Sci. USA* **82**, 6632–6636 (1985).
23. Fiore, S. & Serhan, C.N. Formation of lipoxins and leukotrienes during receptor-mediated interactions of human platelets and recombinant human granulocyte/macrophage colony-stimulating factor-primed neutrophils. *J. Exp. Med.* **172**, 1451–1457 (1990).
24. Ford-Hutchinson, A.W. Leukotriene B4 in inflammation. *Crit. Rev. Immunol.* **10**, 1–12 (1990).
25. Samuelsson, B. Leukotrienes: mediators of immediate hypersensitivity reactions and inflammation. *Science* **220**, 568–575 (1983).
26. Burke, J.A., Levi, R., Guo, Z.G. & Corey, E.J. Leukotrienes C4, D4 and E4: effects on human and guinea-pig cardiac preparations in vitro. *J. Pharmacol. Exp. Ther.* **221**, 235–241 (1982).
27. Roth, D.M. & Lefer, A.M. Studies on the mechanism of leukotriene induced coronary artery constriction. *Prostaglandins* **26**, 573–581 (1983).
28. Wargovich, T., Mehta, J., Nichols, W.W., Pepine, C.J. & Conti, C.R. Reduction in blood flow in normal and narrowed coronary arteries of dogs by leukotriene C4. *J. Am. Coll. Cardiol.* **6**, 1047–1051 (1985).
29. Falk, E., Shah, P.K. & Fuster, V. Coronary plaque disruption. *Circulation* **92**, 657–671 (1995).
30. Dwyer, J.H. *et al.* Arachidonate 5-Lipoxygenase Promoter Genotype, Dietary Arachidonic Acid, and Atherosclerosis. *N. Engl. J. Med.* **350**, 29–37 (2004).
31. Aiello, R.J. *et al.* Leukotriene B4 receptor antagonism reduces monocytic foam cells in mice. *Arterioscler. Thromb. Vasc. Biol.* **22**, 443–449 (2002).
32. Gretarsdottir, S. *et al.* The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat. Genet.* **35**, 131–138 (2003).
33. Gretarsdottir, S. *et al.* Localization of a susceptibility gene for common forms of stroke to 5q12. *Am. J. Hum. Genet.* **70**, 593–603 (2002).
34. Gudmundsson, G. *et al.* Localization of a gene for peripheral arterial occlusive disease to chromosome 1p31. *Am. J. Hum. Genet.* **70**, 586–592 (2002).
35. Gulcher, J.R., Kristjansson, K., Gudbjartsson, H. & Stefansson, K. Protection of privacy by third-party encryption in genetic research in Iceland. *Eur. J. Hum. Genet.* **8**, 739–742 (2000).
36. Kong, A. & Cox, N.J. Allele-sharing models: LOD scores and accurate linkage tests. *Am. J. Hum. Genet.* **61**, 1179–1188 (1997).
37. Gudbjartsson, D.F., Jonasson, K., Frigge, M.L. & Kong, A. Allegro, a new computer program for multipoint linkage analysis. *Nat. Genet.* **25**, 12–13 (2000).
38. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
39. Whittemore, A.S. & Halpern, J. A class of tests for linkage using affected pedigree members. *Biometrics* **50**, 118–127 (1994).
40. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**, 1347–1363 (1996).
41. Nicolae, D. *Allele Sharing Models in Gene Mapping: A Likelihood Approach* (University of Chicago, 1999).
42. Dempster, A., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**, 1–38 (1977).
43. Terwilliger, J.D. & Ott, J. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum. Hered.* **42**, 337–346 (1992).
44. Falk, C.T. & Rubinstein, P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* **51** Pt 3, 227–233 (1987).
45. Lewontin, R.C. The interaction of selection and linkage. ii. Optimum models. *Genetics* **50**, 757–782 (1964).
46. Hill, W.G. & Robertson, A. The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60**, 615–628 (1968).
47. Chen, X., Zehnbauer, B., Gnirke, A. & Kwok, P.Y. Fluorescence energy transfer detection as a homogeneous DNA diagnostic method. *Proc. Natl. Acad. Sci. USA* **94**, 10756–10761 (1997).
48. Steeds, R., Adams, M., Smith, P., Channer, K. & Samani, N.J. Distribution of tissue plasminogen activator insertion/deletion polymorphism in myocardial infarction and control subjects. *Thromb. Haemost.* **79**, 980–984 (1998).
49. Brouilette, S., Singh, R.K., Thompson, J.R., Goodall, A.H. & Samani, N.J. White cell telomere length and risk of premature myocardial infarction. *Arterioscler. Thromb. Vasc. Biol.* **23**, 842–846 (2003).
50. Nomenclature and criteria for diagnosis of ischemic heart disease. Report of the Joint International Society and Federation of Cardiology/World Health Organization task force on standardization of clinical nomenclature. *Circulation* **59**, 607–609 (1979).

Report

Association between the Gene Encoding 5-Lipoxygenase–Activating Protein and Stroke Replicated in a Scottish Population

A. Helgadottir,[1] S. Gretarsdottir,[1] D. St. Clair,[2] A. Manolescu,[1] J. Cheung,[2] G. Thorleifsson,[1] A. Pasdar,[2] S. F. A. Grant,[1] L. J. Whalley,[2] H. Hakonarson,[1] U. Thorsteinsdottir,[1] A. Kong,[1] J. Gulcher[1] K. Stefansson,[1] and M. J. MacLeod[2]

[1]deCODE Genetics, Reykjavik; and [2]Aberdeen Royal Infirmary and University of Aberdeen Medical School, Aberdeen, Scotland

Cardiovascular diseases, including myocardial infarction (MI) and stroke, most often occur on the background of atherosclerosis, a condition attributed to the interactions between multiple genetic and environmental risk factors. We recently reported a linkage and association study of MI and stroke that yielded a genetic variant, HapA, in the gene encoding 5-lipoxygenase–activating protein (*ALOX5AP*), that associates with both diseases in Iceland. We also described another *ALOX5AP* variant, HapB, that associates with MI in England. To further assess the contribution of the *ALOX5AP* variants to cardiovascular diseases in a population outside Iceland, we genotyped seven single-nucleotide polymorphisms that define both HapA and HapB from 450 patients with ischemic stroke and 710 controls from Aberdeenshire, Scotland. The Icelandic at-risk haplotype, HapA, had significantly greater frequency in Scottish patients than in controls. The carrier frequency in patients and controls was 33.4% and 26.4%, respectively, which resulted in a relative risk of 1.36, under the assumption of a multiplicative model (P = .007). We did not detect association between HapB and ischemic stroke in the Scottish cohort. However, we observed that HapB was overrepresented in male patients. This replication of haplotype association with stroke in a population outside Iceland further supports a role for *ALOX5AP* in cardiovascular diseases.

Cardiovascular diseases (CVDs), such as coronary heart disease and stroke, are major causes of death and disability in western societies (Aboderin et al. 2002). As a result of the increasing age of the population, the prevalence of CVD is rising worldwide (American Heart Association 2002). CVDs are largely attributed to atherosclerosis, which has various environmental and genetic risk factors. It is a commonly held view that chronic inflammation initiates and promotes the development of atherosclerotic lesions (Lusis 2000; Libby 2002). Large epidemiologic studies have demonstrated correlations between increased production of markers of systemic inflammation and future cardiovascular events, including myocardial infarction (MI) (Ridker et al. 1997, 1998; Danesh et al. 2000) and stroke (Di Napoli et al. 2001), which supports a central role for inflammation in CVD.

We recently published the association of a variant in the gene encoding 5-lipoxygenase–activating protein (*ALOX5AP* [MIM 603700]) with both MI and stroke in an Icelandic population (Helgadottir et al. 2004). *ALOX5AP*, which encodes an important component of the leukotriene pathway, was identified through a genomewide linkage scan conducted on 296 families with MI and subsequent analysis that determined association with markers within the mapped region on chromosome 13q12-13. A haplotype spanning *ALOX5AP*, HapA, defined by four SNPs, was shown to be associated with MI (relative risk = 1.8; P = .0000023) and, subsequently, the same variant was found to confer risk of stroke in Iceland (relative risk [RR] = 1.7; P = .000095) (Helgadottir et al. 2004). Another SNP-based haplotype within *ALOX5AP*, HapB, showed significant association with MI in British cohorts from Leicester and Sheffield (RR = 2.0; P = .00037) (Helgadottir et al. 2004). We further demonstrated that leukotriene B4 (LTB4) synthesis by neutrophils from patients with a history of MI

Received September 28, 2004; accepted for publication December 13, 2004; electronically published January 7, 2005.

Address for correspondence and reprints: Dr. K. Stefansson, deCODE Genetics Inc., Sturlugata 8, 101 Reykjavik, Iceland. E-mail: kstefans@decode.is

0002-9297/2005/7603-00XX$15.00

is greater than the synthesis by those from controls without MI (Helgadottir et al. 2004).

In the present study, we attempted to replicate the association of *ALOX5AP* with stroke in a population outside Iceland. The SNPs defining HapA (*SG13S25*, *SG13S114*, *SG13S89*, and *SG13S32*) and HapB (*SG13S377*, *SG13S114*, *SG13S41*, and *SG13S35*) were genotyped for 450 Scottish patients who had experienced a stroke and for 710 controls. The patient and control cohorts have been described elsewhere (MacLeod et al. 1999; Meiklejohn et al. 2001; Duthie et al. 2002; Whalley et al. 2004). In brief, 450 patients from northeastern Scotland with CT confirmation of ischemic stroke (including 26 patients with transient ischemic attack [TIA]) were recruited between 1997 and 1999, within 1 wk of admission to the Acute Stroke Unit at Aberdeen Royal Infirmary. Patients were further subclassified in accordance with the TOAST (Trial of Org 10172 in Acute Stroke Treatment) research criteria (Adams et al. 1993). Of the patients, 155 (34.4%) had large-vessel stroke, 96 (21.3%) had cardiogenic stroke, and 109 (24.2%) had small-vessel stroke; for 5 (1.1%) of the patients, stroke with other determined etiology was diagnosed, 7 (1.6%) had more than one etiology, and 78 (17.3%) had unknown cause of stroke despite extensive evaluation. A total of 710 control individuals with no history of stroke or TIA were recruited during follow-up of the 1921 (n = 227) and 1936 (n = 371) Aberdeen Birth Cohort Studies originally recruited in 1932 and 1947, respectively, as part of the Scottish mental surveys (Deary et al. 2004). A further 112 controls were recruited from local primary-care practices (Meiklejohn et al. 2001). Basic clinical characteristics of patients and control individuals are shown in table 1. Approval for the study was granted by the local research ethics committee, and all study participants gave written informed consent.

The haplotype analysis was performed using the program NEMO (Gretarsdottir et al. 2003). NEMO handles missing genotypes and uncertainty with phase through a likelihood procedure, by use of the expectation-maximization algorithm as a computational tool to estimate haplotype frequencies. Since we were testing only two haplotypes, which had been shown elsewhere to confer risk of MI and stroke in an Icelandic cohort and MI in an English cohort, the reported P values are one sided. For the at-risk haplotypes, we calculated RR and population-attributable risk (PAR) under the assumption of a multiplicative model (Falk and Rubinstein 1987; Terwilliger and Ott 1992) in which the risk of the two alleles of haplotypes a person carries multiplies.

The results of the haplotype-association analysis for HapA and HapB are shown in table 2. The haplotype frequencies of HapA in the Scottish populations (patient and control) were higher than in the corresponding Icelandic populations (table 2). As demonstrated in the Icelandic population, the estimated frequency of HapA was significantly greater in Scottish patients who have suffered a stroke than in Scottish controls. The carrier frequency of HapA in Scottish patients and controls was 33.4% and 26.4%, respectively, which resulted in an RR of 1.36 (P = .007) and a corresponding PAR of 9.6%. We had previously observed in the Icelandic population a higher frequency of HapA in male than in female patients with either stroke or MI (Helgadottir et al. 2004). This sex difference in the frequency of HapA was not observed in the Scottish population (table 2).

Table 1

Clinical Characteristics of Scottish Patients and Control Individuals

Characteristics	Patients (n = 450)	Controls (n = 710)
Female:male	42:58	49:51
Age (years)	66.8 ± .6	67.2 ± .4
Hypertension (%)	55.5	23.9
Diabetes (%)	12.6	2.1
Total cholesterol (mmol/liter)	5.65 ± .06	5.64 ± .05

NOTE.—Patients and control individuals were classified as having hypertension and/or diabetes on the basis of previous history or receipt of antihypertensive or antidiabetic therapy. Values with plus-minus symbol (±) are mean ± SE.

We then tested the association of HapB with stroke in the Scottish cohort. HapB has been shown elsewhere to confer risk of MI in an English cohort (Helgadottir et al. 2004). A slight excess of HapB was observed in the patient group (6.8%) compared with controls (5.8%), but it was not significant (table 2). However, sex-specific analysis showed that the frequency of HapB was higher in males with ischemic stroke (9.2%) than in controls, resulting in an RR of 1.65 (P = .016). The frequency of HapB in females with ischemic stroke was 3.5%, which was lower but not significantly different from that of controls. The frequencies of HapB in males and females with ischemic stroke differed significantly (P = .0021). Interestingly, as shown in table 2, similar trends were observed in our Icelandic cohort; the frequency of HapB was greater in males with ischemic stroke (8.6%) than in females with ischemic stroke (5.8%), although this was not significant (P = .055).

To summarize our results, we demonstrate in the present study that HapA, the risk haplotype of *ALOX5AP*, reported elsewhere to confer risk of MI and stroke in an Icelandic cohort, associates with ischemic stroke in a Scottish cohort. HapB, which confers risk of MI in an English cohort, was not associated with ischemic stroke in the Scottish cohort. However, we observed that HapB was overrepresented in male patients.

Historical and archaeological data have suggested a Gaelic ancestry for both Icelanders and Scots. This is

Table 2

Analysis of Association of HapA and HapB with Ischemic Stroke

	HapA			HapB		
Location and Study Population (*n*)	Frequency	RR	*P*	Frequency	RR	*P*
Scotland:						
Controls (710)	.142			.058		
Patients with ischemic stroke (450[a]):	.184	1.36	.007	.068	1.20	NS
Males (253)	.183	1.35	.023	.092	1.65	.016
Females (181)	.179	1.34	.044	.035	.58	NS
Iceland:						
Controls (624)	.095			.067		
Patients with ischemic stroke (632):	.147	1.63	.00013	.073	1.09	NS
Males (335)	.155	1.75	.0002	.086	1.31	NS
Females (297)	.138	1.51	.0079	.058	.86	NS

Note.—Shown are HapA and HapB of *ALOX5AP* and the corresponding number of individuals genotyped, the haplotype frequency in the patient and control cohorts, the RR, and the one-sided *P* values. HapA is defined by the SNPs *SG13S25, SG13S114, SG13S89,* and *SG13S32,* with alleles G, T, G, and A, respectively, and HapB is defined by the SNPs *SG13S377, SG13S114, SG13S41,* and *SG13S35,* with alleles A, A, A, and G, respectively. For SNP genotyping, we used TaqMan assays (Applied Biosystems) or the fluorescent-polarization template-directed dye-terminator incorporation (the SNP-FP-TDI assay), as described elsewhere (Chen et al. 1999). SNP information can be found in the dbSNP database. The DNA used for the SNP genotyping was the product of whole-genome amplification, by use of the GenomiPhi Amplification kit (Amersham), of DNA isolated from the peripheral blood of the Scottish controls and patients with stroke. Data on the Icelandic cohort have been reported elsewhere (Helgadottir et al. 2004). NS = not significant.

[a] Sex unknown for 16 patients.

further supported by recent studies of mtDNA and Y-chromosome diallelic and microsatellite variation in Icelanders, Scandinavians, and Gaels from Ireland and Scotland (Helgason et al. 2000, 2001). Given this common ancestry, it is possible that the two populations share a disease-causing variant and that this variant may reside on the same common haplotype background (HapA). Such a scenario would be consistent with our results; although the estimated RR for HapA in the Scottish cohort is somewhat lower than in the Icelandic cohort, this difference is not statistically significant. Indeed, a similar observation has been made in previous studies of schizophrenia in Iceland and Scotland (Stefansson et al. 2003), in which the same extended haplotype was found to confer risk of schizophrenia in both populations, with comparable frequencies in patient and control groups in the two countries.

The gene *ALOX5AP* encodes the membrane-associated 5-lipoxygenase–activating protein (FLAP), an important mediator of the activity of cellular 5-lipoxygenase (5-LO), which is a key enzyme in the biosynthesis of leukotrienes (Dixon et al. 1990; Miller et al. 1990). Leukotrienes are proinflammatory mediators produced predominantly in inflammatory cells such as polymorphonuclear leukocytes, macrophages, and mast cells. Over the last decade, a number of studies have supported an important role for inflammation in atherosclerosis—from atheroma initiation to promotion of plaque rupture, thereby triggering thrombosis, the main atherosclerotic complication that causes MI and stroke (Libby 2002). The 5-LO pathway could be an important contributor to the pathophysiology of atherosclerosis through the formation of the proinflammatory LTB4 and/or through an increase in vascular permeability caused by cysteinyl leukotrienes. Indeed, we have shown increased production of LTB4 in neutrophils from patients with history of MI, compared with controls without history of MI (Helgadottir et al. 2004). This is further supported by recent human-expression studies (Spanbroek et al. 2003) that show an increased expression of members of the 5-LO pathway, including 5-LO and FLAP, in atherosclerotic lesions at various stages of their development. Moreover, a promoter variant of 5-LO (*ALOX5* [MIM 152390]) has been shown to be associated with increased carotid artery intima-media thickness and with heightened inflammatory biomarkers (Dwyer et al. 2004). In addition, an atherosclerotic mouse model with a heterozygous deficiency of 5-LO shows resistance to atherosclerosis (Mehrabian et al. 2002), and an LTB4 receptor antagonist blocks the development of atherosclerosis in apoE- and LDLR-deficient mice (Aiello et al. 2002; Mehrabian et al. 2002). Together, these studies suggest that chronic upregulation of the leukotriene pathway may be harmful to the vasculature, in terms of atherosclerosis progression and plaque instability.

The precise mechanism by which the *ALOX5AP* variants confer risk of MI and stroke is still unclear. As reported elsewhere, we have not observed SNPs in the coding sequence that led to amino acid substitution (Helgadottir et al. 2004). Therefore, one can speculate that unidentified variation in regulatory regions of the gene—that affects transcription, splicing, message stability, message transport, or translation efficiency—may underlie the risk conferred by *ALOX5AP*.

The results of the present study show that HapA associates with ischemic stroke in a Scottish population, thereby providing replication of work that showed that the same haplotype confers increased risk of stroke in an Icelandic population. This replication constitutes additional evidence for the role of *ALOX5AP* in the pathogenesis of stroke. Identification of genetic risk factors for the common forms of stroke may facilitate identification of individuals at increased risk and may lead to novel strategies for the prevention and treatment of stroke.

Acknowledgments

The authors thank Thorbjorg Jonsdottir and personnel at the deCODE core facilities for valuable contributions. The authors acknowledge the support of the Chief Scientist's Office, Scottish Executive.

Electronic-Database Information

The URLs for data presented herein are as follows:

dbSNP, http://www.ncbi.nlm.nih.gov/SNP/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *ALOX5AP* and *ALOX5*)

References

Aboderin I, Kalache A, Ben-Shlomo Y, Lynch JW, Yajnik CS, Kuh D, Yach D (2002) Life course perspectives on coronary heart disease, stroke and diabetes: key issues and implications for policy and research. World Health Organization, Geneva

Adams HP Jr, Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL, Marsh EE 3rd (1993) Classification of subtype of acute ischemic stroke: definitions for use in a multicenter clinical trial. Stroke 24:35–41

Aiello RJ, Bourassa PA, Lindsey S, Weng W, Freeman A, Showell HJ (2002) Leukotriene B4 receptor antagonism reduces monocytic foam cells in mice. Arterioscler Thromb Vasc Biol 22:443–449

American Heart Association (2002) Heart disease and stroke statistics: 2003 update, Dallas

Chen X, Levine L, Kwok PY (1999) Fluorescence polarization in homogeneous nucleic acid analysis. Genome Res 9:492–498

Danesh J, Whincup P, Walker M, Lennon L, Thomson A, Appleby P, Gallimore JR, Pepys MB (2000) Low grade inflammation and coronary heart disease: prospective study and updated meta-analyses. BMJ 321:199–204

Deary IJ, Whiteman MC, Starr JM, Whalley LJ, Fox HC (2004) The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947. J Pers Soc Psychol 86:130–147

Di Napoli M, Papa F, Bocola V (2001) C-reactive protein in ischemic stroke: an independent prognostic factor. Stroke 32: 917–924

Dixon RA, Diehl RE, Opas E, Rands E, Vickers PJ, Evans JF, Gillard JW, Miller DK (1990) Requirement of a 5-lipoxygenase-activating protein for leukotriene synthesis. Nature 343:282–284

Duthie SJ, Whalley LJ, Collins AR, Leaper S, Berger K, Deary IJ (2002) Homocysteine, B vitamin status, and cognitive function in the elderly. Am J Clin Nutr 75:908–913

Dwyer JH, Allayee H, Dwyer KM, Fan J, Wu H, Mar R, Lusis AJ, Mehrabian M (2004) Arachidonate 5-lipoxygenase promoter genotype, dietary arachidonic acid, and atherosclerosis. N Engl J Med 350:29–37

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Gretarsdottir S, Thorleifsson G, Reynisdottir ST, Manolescu A, Jonsdottir S, Jonsdottir T, Gudmundsdottir T, et al (2003) The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. Nat Genet 35:131–138

Helgadottir A, Manolescu A, Thorleifsson G, Gretarsdottir S, Jonsdottir H, Thorsteinsdottir U, Samani NJ, et al (2004) The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. Nat Genet 36:233–239

Helgason A, Hickey E, Goodacre S, Bosnes V, Stefánsson K, Ward R, Sykes B (2001) mtDNA and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. Am J Hum Genet 68:723–737

Helgason A, Sigurðardóttir S, Nicholson J, Sykes B, Hill EW, Bradley DG, Bosnes V, Gulcher JR, Ward R, Stefánsson K (2000) Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland. Am J Hum Genet 67:697–717

Libby P (2002) Inflammation in atherosclerosis. Nature 420: 868–874

Lusis AJ (2000) Atherosclerosis. Nature 407:233–241

MacLeod MJ, Dahiyat MT, Cumming A, Meiklejohn D, Shaw D, St Clair D (1999) No association between glu/asp polymorphism of *NOS3* gene and ischemic stroke. Neurology 53:418–420

Mehrabian M, Allayee H, Wong J, Shi W, Wang XP, Shaposhnik Z, Funk CD, Lusis AJ, Shih W (2002) Identification of 5-lipoxygenase as a major gene contributing to atherosclerosis susceptibility in mice. Circ Res 91:120–126

Meiklejohn DJ, Vickers MA, Dijkhuisen R, Greaves M (2001) Plasma homocysteine concentrations in the acute and convalescent periods of atherothrombotic stroke. Stroke 32:57–62

Miller DK, Gillard JW, Vickers PJ, Sadowski S, Léveillé C, Mancini JA, Charleson P, Dixon RAF, Ford-Hutchinson AW, Fortin R, Gautier JY, Rodkey J, Rosen R, Rouzer C, Sigal IS, Strader CD, Evans JF (1990) Identification and isolation of a membrane protein necessary for leukotriene production. Nature 343:278–281

Ridker PM, Buring JE, Shih J, Matias M, Hennekens CH (1998) Prospective study of C-reactive protein and the risk of future cardiovascular events among apparently healthy women. Circulation 98:731–733

Ridker PM, Cushman M, Stampfer MJ, Tracy RP, Hennekens CH (1997) Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. N Engl J Med 336:973–979

Spanbroek R, Grabner R, Lotzer K, Hildner M, Urbach A, Ruhling K, Moos MP, Kaiser B, Cohnert TU, Wahlers T, Zieske A, Plenz G, Robenek H, Salbach P, Kuhn H, Radmark O, Samuelsson B, Habenicht AJ (2003) Expanding expression of the 5-lipoxygenase pathway within the arterial wall during human atherogenesis. Proc Natl Acad Sci USA 100: 1238–1243

Stefansson H, Sarginson J, Kong A, Yates P, Steinthorsdottir V, Gudfinnsson E, Gunnarsdottir S, Walker N, Petursson H, Crombie C, Ingason A, Gulcher JR, Stefansson K, St Clair D (2003) Association of neuregulin 1 with schizophrenia confirmed in a Scottish population. Am J Hum Genet 72: 83–87

Terwilliger JD, Ott J (1992) A haplotype-based "haplotype relative risk" approach to detecting allelic associations. Hum Hered 42:337–346

Whalley LJ, Fox HC, Wahle KW, Starr JM, Deary IJ (2004) Cognitive aging, childhood intelligence, and the use of food supplements: possible involvement of n-3 fatty acids. Am J Clin Nutr 80:1650–1657

European Journal of Human Genetics (2007) 15, 959–966
© 2007 Nature Publishing Group All rights reserved 1018-4813/07 $30.00
www.nature.com/ejhg

ARTICLE

ALOX5AP gene variants and risk of coronary artery disease: an angiography-based study

Domenico Girelli*[,1], Nicola Martinelli[1], Elisabetta Trabetti[2], Oliviero Olivieri[1], Ugo Cavallari[2], Giovanni Malerba[2], Fabiana Busti[1], Simonetta Friso[1], Francesca Pizzolo[1], Pier Franco Pignatti[2] and Roberto Corrocher[1]

[1]*Department of Clinical and Experimental Medicine, University of Verona, Verona, Italy;* [2]*Department of Mother and Child and Biology-Genetics, University of Verona, Verona, Italy*

The aim of this study was to explore the role of variants of the gene encoding arachidonate 5-lipoxygenase-activating protein (*ALOX5AP*) as possible susceptibility factors for coronary artery disease (CAD) and myocardial infarction (MI) in patients with or without angiographically proven CAD. A total of 1431 patients with or without angiographically documented CAD were examined simultaneously for seven *ALOX5AP* single-nucleotide polymorphisms, allowing reconstruction of the at-risk haplotypes (HapA and HapB) previously identified in the Icelandic and British populations. Using a haplotype-based approach, HapA was not associated with either CAD or MI. On the other hand, HapB and another haplotype within the same region (that we named HapC) were significantly more represented in CAD *versus* CAD-free patients, and these associations remained significant after adjustment for traditional cardiovascular risk factors by logistic regression (HapB: odds ratio (OR) 1.67, 95% confidence interval (CI) 1.04–2.67; $P=0.032$; HapC: OR 2.41, 95% CI 1.09–5.32; $P=0.030$). No difference in haplotype distributions was observed between CAD subjects with or without a previously documented MI. Our angiography-based study suggests a possible modest role of *ALOX5AP* in the development of the atheroma rather than in its late thrombotic complications such as MI.

European Journal of Human Genetics (2007) 15, 959–966; doi:10.1038/sj.ejhg.5201854; published online 16 May 2007

Keywords: ALOX5AP; coronary artery disease; myocardial infarction

Introduction

Interest in unraveling the genetic basis of coronary artery disease (CAD) has been recently renewed by results obtained applying powerful approaches such as genome-wide scan studies.[1–3] At variance with classic association studies involving single-nucleotide polymorphisms (SNPs) in candidate genes, genome-wide scan studies have the advantage of discovering new gene(s), without *a priori* hypothesis. A paradigm of the successful use of such strategies was the identification of arachidonate 5-lipoxygenase-activating protein (*ALOX5AP*) as a susceptibility gene for myocardial infarction (MI) and stroke.[4] Interestingly, *ALOX5AP* encodes the 5-lipoxygenase-activating protein (FLAP), which is an essential regulator of the biosynthesis of the leukotriene A4 (LTA4).[5,6] Indeed, the 5-lipoxygenase (5-LO)/leukotriene pathway has been independently implicated in the pathogenesis of atherosclerosis in humans[7,8] and mice[9] (reviewed by Zhao and Funk[10]). While not successful in discovering causal variants in *ALOX5AP*, the original study by Helgadottir *et al*[4] identified a 4-SNP haplotype, named HapA, as a risk factor for MI and stroke in the Icelandic population. The Authors were unable to confirm the result in a cohort of British patients

*Correspondence: Professor D Girelli, Department of Clinical and Experimental Medicine, University of Verona, Policlinico GB Rossi, 37134 Verona, Italy. Tel: +39 45 8124403; Fax: +39 45 580111; E-mail: domenico.girelli@univr.it

Received 27 October 2006; revised 9 March 2007; accepted 17 April 2007; published online 16 May 2007

with MI; however, in such cohort they reported an association of another 4-SNP haplotype, named HapB, with MI risk.[4] Few subsequent studies on *ALOX5AP* in different populations yielded conflicting results. Löhmussaar *et al*[11] studied Central European patients with stroke, finding a significant association for several *ALOX5AP* SNPs, including one that was part of HapA. On the other hand, studies in North Americans failed to show a significant association with either stroke or MI.[12,13]

To date, no genetic–epidemiological data are available for populations from Southern Europe. Moreover, none of the previous studies specifically attempted to dissect the role of *ALOX5AP* in the atherosclerosis phenotype rather than in its 'complication' phenotype (MI). We therefore evaluated simultaneously seven *ALOXA5* SNPs and their reconstructed haplotypes as possible risk determinants for CAD and MI within the framework of an Italian population with or without angiographically confirmed CAD.

Materials and methods

Study population

The Verona Heart Project is an ongoing study aimed to identify new risk factors for CAD and MI in a population of subjects with angiographic documentation of their coronary vessels. Details about the enrolment criteria have been described elsewhere.[14] In the present study, we examined data from a total of 1431 subjects, for whom complete analyses of seven *ALOX5AP* SNPs (see below) were available. Of these subjects, 1047 had angiographically documented severe coronary atherosclerosis (CAD group), the majority of them being candidates for coronary artery bypass grafting or percutaneous coronary intervention. The disease severity was evaluated by counting the number of major epicardial coronary arteries (left anterior descending, circumflex, and right) affected with ≥ 1 significant stenosis ($\geq 50\%$). On the other hand, 384 subjects had completely normal coronary arteries, being submitted to coronary angiography for reasons other than CAD, mainly valvular heart disease (CAD-free group). Controls were also required to have neither history nor clinical or instrumental evidence of atherosclerosis in vascular districts beyond the coronary bed. Since the primary aim of our selection was to provide an objective and clear-cut definition of the atherosclerotic phenotype, subjects with nonsignificant coronary stenosis (ie $<50\%$) were not included in the study. CAD subjects were classified into MI and non-MI subgroups by combining data from history with a thorough review of medical records showing diagnostic electrocardiogram and enzyme changes, and/or the typical sequelae of MI on ventricular angiography. An appropriate documentation was obtained for 1046/1047 (99.9%) CAD patients: from those 624 subjects had a history of previous MI, whereas the remaining 422 subjects had no history of MI. The angiograms were assessed by two cardiologists unaware that the patients were to be included in the study. Samples of venous blood were drawn from each subject after an overnight fast. Serum lipids and the other routine biochemical parameters were determined as described previously.[14] At the time of blood sampling, a complete clinical history was collected, including the assessment of cardiovascular risk factors such as obesity, smoking, hypertension, and diabetes.

The study was approved by our local Ethical Committee. Informed consent was obtained from all the patients after a full explanation of the study.

Genotyping

To make possible comparison with studies in other populations, we selected seven previously described *ALOX5AP* (GeneID: 241; chromosome: 13q12) SNPs (SG13S25, SG13S377, SG13S114, SG13S89, SG13S32, SG13S41 and SG13S35), maintaining their original nomenclature,[4] as well as the nomenclature of the reconstructed haplotypes. The seven SNPs were initially tested by PCR and restriction analyses (Supplementary Table 1) in a small group of randomly chosen DNA samples in order to verify the heterozygosity in the study population. All the samples were then genotyped in two multiplex reactions for six SNPs (SG13S377, SG13S41, SG13S32, and SG13S114 in multiplex one, M1; SG13S25, SG13S35 in multiplex two, M2) using LightCycler™ real-time PCR technology based on fluorescence resonance energy transfer and melting point analysis. The sequences of primers and probes used for the six SNPs genotyping with melting point analysis are shown in Supplementary Table 2. Both primers and fluorescently labelled probes were synthesized by Sigma-Proligo (Proligo France SAS). PCR and melting curve analysis was performed in $20\,\mu l$ volumes in glass capillaries (Hoffmann-La Roche). PCR conditions for M1 and M2 are detailed in Supplementary Tables 3 and 4, respectively. Cycling and melting curve analysis conditions were different for the two multiplex reactions, as given in Supplementary Table 5. As the SG13S89 polymorphism was not easily detectable in a multiplex reaction, it was genotyped by PCR and restriction analysis for all the samples, using the following primers forward (F): 5'-AAGTGCATCTCAAGGAGGT-3' and reverse (R) 5'-ATTAG CAGAAGAGCCAAGT-3'.

Statistical analysis

The analyses were performed mainly with SSPS 13.0 statistical package (SPSS Inc., Chicago, IL, USA). Distributions of continuous variables in groups were expressed as means±SD. Quantitative data were assessed using the Student's *t*-test. Associations between qualitative variables were analysed with the χ^2 test or Fisher exact test, when indicated. Allele and genotype frequencies among cases and controls were compared with values predicted by

Hardy–Weinberg equilibrium using χ^2 test. To assess the association with CAD or MI, relative risks associated with each genotype were calculated separately by univariate logistic regression and then by multiple logistic regression adjusted for the traditional cardiovascular risk factors (ie sex, age, smoking, hypertension, diabetes, total cholesterol, and triglycerides), assuming an additive, dominant or recessive mode of inheritance.

Pairwise linkage disequilibrium was examined as described by Devlin and Risch.[15] Haplotype frequencies were estimated using R software with haplo.stats package (R Foundation for Statistical Computing, Vienna, Austria; ISBN 3-900051-07-0, URL: http://www.R-project.org).[16] The upper and lower bounds of the 95% confidence interval (CI) were calculated by simulating 1000 random samples from a population having the haplotype frequencies estimated on the entire sample set. The *D′* measure was calculated for each simulated sample. The upper and lower bounds represent the quantiles corresponding to the 0.025 or 0.975 probabilities of the *D′* distribution. Haplotype blocks were defined as proposed by Gabriel *et al.*[17] The relationship between haplotypes and clinical outcomes was examined using a generalized linear model regression of a trait on haplotype effects, allowing for ambiguous haplotypes (haplo.glm function),[16] and adjusting for the above-mentioned risk factors. Randomization test by permuting the cases and controls was performed by means of Monte Carlo method to confirm the results. Haplotypes present in less than 10 individuals were not considered in the analyses.

The study power was assessed by means of the Altman nomogram, after adjustment for the asymmetric distribution of population subgroups (CAD-free *versus* CAD; non-MI *versus* MI). The study has adequate power (>90%) to replicate the findings for odds ratios (ORs) greater than 2.0, which is consistent with those observed in the previous studies.[4] For each OR, 95% CIs were calculated. A value of two-tailed $P<0.05$ was considered significant.

Results

Table 1a summarizes the clinical characteristics of the study population stratified according to the presence (CAD) or absence (CAD-free) of angiographically documented CAD. As expected, traditional cardiovascular risk factors were more represented in the CAD group. The characteristics of the CAD population, divided in two subgroups according to the presence or absence of a previous documented MI, are reported in Table 1b. The genotype frequencies for the polymorphisms tested were in Hardy–Weinberg equilibrium both in the CAD and CAD-free groups.

Allele and genotype distributions were similar either between CAD and CAD-free groups (Table 2a) or within CAD subjects with or without a previous MI (Table 2b). Results from the regression analyses, assuming additive, dominant or recessive mode of inheritance, showed no significant association of the gene variants tested with the clinical outcomes (data not shown). In general, the SNPs tested were in linkage disequilibrium, as shown in Table 3.

Considering haplotype analysis, the most frequent haplotypes were G-T-G-C and G-T-A-G for HapA SNPs and HapB/C SNPs, respectively, and thus were used as the referents. The haplotype distributions for HapA SNPs were similar between CAD and CAD-free subjects ($P=0.937$). On the other hand, the haplotype distributions for HapB SNPs were significantly different between CAD and CAD-free subjects ($P=0.014$), as shown in Table 4a. More precisely, two haplotypes A-A-A-G (HapB) and G-T-A-A (that we named HapC) were more represented in CAD group (7.5 *versus* 5.5% and 3.7 *versus* 1.6%, respectively), and these associations remained significant also after adjustment for

Table 1a Clinical characteristics of the study population stratified according to absence (CAD-free) or presence (CAD) of angiographically documented CAD

	CAD-free (n = *384*)	*CAD* (n = *1047*)	P-*values*
Age (years)	58.7±12.3	61.2±9.8	<0.001*
Males (%)	65.6	79.8	<0.001#
BMI (kg/m^2)	25.4±3.5	26.8±3.5	0.936*
Hypertension (%)	40.5	66.3	<0.001#
Smoking (%)	43.9	67.9	<0.001#
Diabetes (%)	6.8	19.2	<0.001#
Total cholesterol (mmol/l)	5.47±1.10	5.54±1.17	0.322*
Triglycerides (mmol/l)	1.49±0.67	1.91±0.99	<0.001*

*By *t*-test; #by χ^2 test.

Table 1b Clinical characteristics of the CAD patients, with (MI) or without (no-MI) a previous documented MI

	No-MI (n = *422*)	*MI* (n = *624*)	P-*values*
Age (years)	62.5±8.9	60.4±10.2	0.002*
Males (%)	75.6	82.7	0.005#
BMI (kg/m^2)	27.0±3.5	26.6±3.6	0.583*
Hypertension (%)	72.3	62.1	0.001#
Smoking (%)	62.3	71.7	0.002#
Diabetes (%)	19.3	19.2	0.592#
Total cholesterol (mmol/l)	5.6±1.1	5.5±1.2	0.677*
Triglycerides (mmol/l)	1.9±0.9	1.9±1.0	0.396*
CAD severity			
One vessel	24.3	12.4	<0.001#
Two vessels	26.0	24.1	
Three vessels	48.1	61.8	
Left main coronary artery	1.7	1.7	

*By *t*-test; #by χ^2 test.

Table 2a *ALOX5AP* genotype and allele distribution in the study population stratified according to absence (CAD-free) or presence (CAD) of angiographically documented CAD

ALOX5AP genotype, %	*CAD-free (n = 384)*	*CAD (n = 1047)*	P-*values**
SG13S25			
GG	85.7	84.6	
GA	13.8	14.8	0.885
AA	0.5	0.6	
G allele	92.6	92.0	0.682
A allele	7.4	8.0	
SG13S377			
GG	76.8	73.9	
GA	20.8	24.6	0.180
AA	2.3	1.4	
G allele	87.2	86.2	0.530
A allele	12.8	13.8	
SG13S114			
TT	42.4	40.6	
TA	41.7	44.8	0.559
AA	15.9	14.6	
T allele	63.3	63.0	0.921
A allele	36.7	37.0	
SG13S89			
GG	86.7	87.5	
GA	12.8	12.1	0.794
AA	0.5	0.4	
G allele	93.1	93.6	0.727
A allele	6.9	6.4	
SG13S32			
AA	22.9	22.6	
AC	51.6	49.9	0.747
CC	25.5	27.5	
C allele	51.3	52.4	0.620
A allele	48.7	47.6	
SG13S41			
AA	81.8	81.9	
AG	17.4	17.3	0.997
GG	0.8	0.8	
A allele	90.5	90.6	0.995
G allele	9.5	9.4	
SG13S35			
GG	83.9	81.5	
GA	15.6	18.1	0.528
AA	0.5	0.4	
G allele	91.7	90.5	0.396
A allele	8.3	9.5	

*By χ^2 test or Fisher's exact test.

Table 2b *ALOX5AP* genotype and allele distribution in the CAD group stratified according to absence (no-MI) or presence (MI) of previously documented MI

ALOX5AP genotype	*No-MI (n = 422)*	*MI (n = 624)*	P-*values**
SG13S25			
GG	84.6	84.6	
GA	14.2	15.2	0.107
AA	1.2	0.2	
G allele	91.7	92.2	0.727
A allele	8.3	7.8	
SG13S377			
GG	72.0	75.2	
GA	26.3	23.6	0.509
AA	1.7	1.3	
G allele	85.2	87.0	0.283
A allele	14.8	13.0	
SG13S114			
TT	37.7	42.6	
TA	47.4	42.9	0.263
AA	14.9	14.4	
T allele	61.4	64.1	0.222
A allele	38.6	35.9	
SG13S89			
GG	87.4	87.7	
GA	12.3	11.9	0.868
AA	0.2	0.5	
G allele	93.6	93.6	0.936
A allele	6.4	6.4	
SG13S32			
AA	23.9	21.8	
AC	49.1	50.3	0.719
CC	27.0	27.9	
C allele	51.5	53.0	0.528
A allele	48.5	47.0	
SG13S41			
AA	83.4	81.1	
AG	16.1	17.9	0.515
GG	0.5	1.0	
A allele	91.5	90.1	0.315
G allele	8.5	9.9	
SG13S35			
GG	81.3	81.7	
GA	18.7	17.6	0.282
AA	0	0.6	
G allele	90.6	90.5	0.997
A allele	9.4	9.5	

*By χ^2 test or Fisher's exact test.

traditional cardiovascular risk factors, that is, sex, age, smoking, hypertension, diabetes, total cholesterol, and triglycerides (Table 4b). The significance of the general model, including genetic factors arranged as haplotypes, was confirmed after randomization test ($P = 0.022$ for general model, $P = 0.013$ for HapB and $P = 0.021$ for HapC, after 1000 permutations).

There was no difference in haplotype distributions between CAD subjects with or without a previous MI, either for HapA region or for HapB region (Table 4c).

Discussion

The present investigation in Italian patients provides some evidence that the *ALOX5AP* gene might play a role in

Table 3 Pairwise linkage disequilibrium analysis

		R²						
		SG13S25	*SG13S377*	*SG13S114*	*SG13S89*	*SG13S32*	*SG13S41*	*SG13S35*
D′	SG13S25	—	0.013	0.045	0.006	0.087	0.007	0.002
	SG13S377	1.000	—	0.185	0.003	0.113	0.007	0.145
	SG13S114	0.956	0.834	—	0.072	0.038	0.105	0.026
	SG13S89	1.000	0.544	0.774	—	0.050	0.463	0.004
	SG13S32	0.969	0.815	0.245	0.882	—	0.093	0.077
	SG13S41	0.920	0.666	0.770	0.828	0.987	—	0.009
	SG13S35	0.529	0.473	0.392	0.790	0.836	0.914	—

D′, Lewontin normalized value; *R*, correlation coefficient.

conferring susceptibility to CAD also in South European populations. Nonetheless, since the statistically significant association we found was relatively weak, the role of this gene, if any, seems modest. To put our results into a more general perspective, we propose the following considerations.

Comparison with previous studies

The landmark study by Helgadottir *et al* identified two different haplotypes as CAD risk factors in populations of different ancestry. According to the haplotype block definition proposed by Gabriel *et al*,[17] we observed three haplotype blocks of two SNPs each (block 1: SG13S25-SG12S377; block 2: SG13S32-SG13S41; block 3: SG13S41-SG13S35). Therefore, the SNPs describing HapA or HapB/C do not define a single haplotype block. This finding is consistent with what observed by Helgadottir *et al.*[4]

In Icelandics (a genetic isolate), the HapA conferred a nearly twofold risk of MI.[4] This was not confirmed in British patients, in whom, on the other hand, a different 4-SNPs haplotype (HapB) was associated with MI. Neither HapA nor HapB was associated to incident MI in a cohort of male US physicians.[13] To make possible a comparison, we focused on a standardized set of seven *ALOX5AP* SNPs, allowing reconstruction of the same at-risk haplotypes reported in the literature. With respect to HapA, our results suggest that this haplotype may not be informative for risk assessment of CAD in non-Icelandic populations. With respect to HapB, a modest contribution of this genetic marker to CAD risk was observed. Haplotype analyses revealed in our population a nominally significant association between CAD and another *ALOX5AP* haplotype ('HapC'), unremarkable in previous studies. Considering also the low frequency of this haplotype, the relevance of this finding remains uncertain. The observed differences among populations are not surprising, and may relate in part to population-specific differences in allele and haplotype frequencies (for a summary of previous studies see Table 5). For example, the frequency of HapA in Icelandic controls (9.5%) was well below that observed in North American (15%), German (15.2%), and our Italian (18.6%) populations. Moreover, it has to be underscored that we are dealing with disease-risk-associated haplotypes made of SNPs with no obvious potential effects on function, whose association(s) with yet unidentified causal variant(s) in *ALOX5AP* may differ between populations with differing genealogies. In other words, it would not be unexpected to find in the future different pathogenic *ALOX5AP* mutation(s), with different frequencies among populations, arising on different haplotype background. Noteworthy, a replication study in a Japanese population[18] found an allele frequency of HapA/HapB SNPs too low to conduct meaningful association. Nevertheless, in that population haplotypes constructed on the basis of two other intronic SNPs were significantly associated with MI.

ALOX5AP, leukotriene pathway, and CAD pathogenesis

Preliminary functional data by Helgadottir *et al* indicated that some at-risk haplotypes were associated to increased neutrophil release of leukotriene B4 (LTB4). Being LTB4 synthesized from LTA4, it implies that *ALOX5AP* variants might determine proinflammatory gain of functions. The role of inflammation in CAD pathogenesis is now well-established (reviewed by Hansson[19]). The FLAP protein encoded by *ALOX5AP* has an important role in the initial steps of the biosynthesis of leukotrienes,[5,6] which in turn have a variety of proinflammatory effects.[20] Besides the *ALOX5AP* story, genetic evidence for the involvement of the 5-LO/leukotriene pathway in CAD is accumulating.[21–23] The same Icelandic group recently reported that another gene involved in this pathway, that is, leukotriene A4 hydrolase, conferred risk of CAD, especially in African Americans.[21] Dwyer *et al*[22] found an association between promoter variations of the *ALOX5* gene (encoding 5-LO, ie the FLAP target) and carotid intima-media thickness (a preclinical surrogate marker of atherosclerosis). As a functional counterpart of intriguing genetic studies, a bulk of animal experiments have linked the 5-LO pathway to atherosclerosis, although results are sometimes discordant (critically reviewed by Funk[23]). Interestingly, many of basic researches leading to the 'lipoxygenases hypothesis'[24,25] points towards an involvement in early events of atheroma development, through LTB4-mediated migration and

Table 4 *ALOX5AP* haplotype distribution in the study population stratified according to the presence or absence of angiographically documented CAD (a); ORs with 95% CIs for CAD for HapB region haplotypes, calculated by means of haplotype-based logistic regression analysis adjusted for traditional risk factors for CAD, that is, sex, age, smoking, hypertension, diabetes, and plasma lipids (b); *ALOX5AP* haplotype distribution in the CAD group stratified according to absence (no-MI) or presence (MI) of previously documented myocardial infarction (c)

(a)

ALOX5AP haplotype	*CAD-free (%)*	*CAD (%)*	*P-values**
HapA SNPs (SG13S25, SG13S114, SG13S89, SG13S32)			
G-T-G-A (HapA)[a]	18.6	16.9	0.937
G-T-G-C	35.8	37.5	
G-A-G-A	22.7	22.3	
G-A-G-C	8.6	8.8	
G-A-A-C	5.5	5.5	
A-T-G-A	7.4	7.8	
HapB/C SNPs (SG13S377, SG13S114, SG13S41, SG13S35)			
G-T-A-G	58.3	56.8	**0.014**
G-T-A-A (HapC)[a]	**1.6**	**3.7**	
G-T-G-G	1.4	1.1	
G-A-A-G	17.1	16.0	
G-A-G-G	7.3	7.7	
A-T-A-G	1.5	1.0	
A-A-A-G (HapB)[a]	**5.5**	**7.5**	
A-A-A-A	4.8	4.6	

(b)

HapB/C SNPs	*OR for CAD*	*P-values#*	*P-values^*
G-T-A-A (HapC)[a]	**2.41 (1.09–5.32)**	**0.030**	**0.021**
A-A-A-G (HapB)[a]	**1.67 (1.04–2.67)**	**0.032**	**0.013**

(c)

ALOX5AP haplotype	*No-MI (%)*	*MI (%)*	*P-values**
HapA SNPs (SG13S25, SG13S114, SG13S89, SG13S32)			
G-T-G-A (HapA)[a]	15.7	17.6	0.587
G-T-G-C	36.6	38.2	
G-A-G-A	24.1	21.2	
G-A-G-C	8.9	8.8	
G-A-A-C	5.7	5.4	
A-T-G-A	8.3	7.5	
HapB/C SNPs (SG13S377, SG13S114, SG13S41, SG13S35)			
G-T-A-G	55.6	57.5	0.547
G-T-A-A (HapC)[a]	3.5	3.9	
G-T-G-G	0.9	1.3	
G-A-A-G	17.2	15.2	
G-A-G-G	6.9	8.3	
A-T-A-G	1.0	1.1	
A-A-A-G (HapB)[a]	8.4	7.0	
A-A-A-A	4.5	4.7	

*By regression analysis.

#By regression analysis adjusted for sex, age, smoking, hypertension, diabetes, total cholesterol, and triglycerides.

^By randomization test after 1000 permutations.

[a] HapA is defined by SG13S25, SG13S114, SG13S89, and SG13S32 SNPs, with alleles G, T, G, A, respectively. HapB and C are defined by SG13S377, SG13S114, SG13S41, and SG13S35 SNPs, with alleles A, A, A, G, or G, T, A, A, respectively.

Bold characters underscore the haplotypes with a significant different distribution.

Table 5 Frequencies of HapA, HapB, and some *ALOX5AP* SNPs in studies published so far (in patients with MI or stroke) and in our study

Study (author's name)	*Controls (%)*	*Cases (%)*	*P-values*
Helgadottir et al,[4]			
Icelandic cohort			
HapA	9.5	15.8[a], 14.9[b]	*<0.001*[a], *<0.001*[b]
SG13S114 *allele T*	65.8	70.0[a]	*0.021*
British cohort			
HapA	16.8	15.1[a]	*NS*
HapB	4.0	7.5[a]	*<0.001*
Löhmussaar et al,[11]			
German cohort			
HapA	15.2	14.5[b]	*NS*
SG13S25 *allele G*	90.1	89.4[b]	*NS*
SG13S114 *allele T*	65.0	68.5[b]	*0.025*
SG13S89 *allele G*	96.0	94.7[b]	*NS*
SG13S32 *allele A*	46.7	46.9[b]	*NS*
Helgadottir et al,[29]			
Scottish cohort			
HapA	14.2	18.4[b]	*0.007*
HapB	5.8	6.8[b]	*NS*
Kajimoto et al,[18]			
Japanese cohort			
SG13S25 *allele G*	99.97	100[a]	*0.557*
SG13S377 *allele G*	81.6	80.0[a]	*0.243*
SG13S114 *allele T*	64.7	64.1[a]	*0.298*
SG13S89 *allele G*	99.2	99.0[a]	*0.603*
SG13S32 *allele A*	64.9	65.1[a]	*0.428*
SG13S41 *allele A*	99.2	98.7[a]	*0.303*
SG13S35 *allele G*	100	100[a]	—
A162C allele C	48.8	44.7[a]	*0.129*
T8733A *allele A*	43.6	42.6[a]	*0.570*
Haplotype 162A-8733A	20.0	25.8[a]	*0.003*
Haplotype 162C-8733A	23.6	16.9[a]	*0.001*
Meschia et al,[12]			
North American cohort			
SG13S25 *allele G*	87.9	89.7[b]	*0.200*
SG13S114 *allele T*	57.8	59.1[b]	*0.180*
SG13S89 *allele G*	87.4	91.2[b]	*0.150*
SG13S32 *allele A*	49.2	51.3[b]	*0.790*
Zee et al,[13]			
US cohort			
HapA	14[c], 15[d]	17[a], 18[b]	*0.460*[a], *0.710*[b]
HapB	7[c], 7[d]	6[a], 8[b]	*0.080*[a], *0.470*[b]
SG13S25 *allele G*	90[c], 1[d]	90[a], 90[b]	*0.890*[a], *0.470*[b]
SG13S377 *allele G*	87[c], 83[d]	88[a], 87[b]	*0.410*[a], *0.150*[b]
SG13S114 *allele T*	68[c], 63[d]	68[a], 63[b]	*0.630*[a], *0.990*[b]
SG13S89 *allele G*	95[c], 94[d]	94[a], 94[b]	*0.840*[a], *0.960*[b]
SG13S32 *allele A*	46[c], 52[d]	50[a], 52[b]	*0.150*[a], *0.990*[b]
SG13S41 *allele A*	91[c], 92[d]	91[a], 92[b]	*0.730*[a], *0.680*[b]
SG13S35 *allele G*	91[c], 89[d]	93[a], 91[b]	*0.210*[a], *0.260*[b]
This study, 2006			
Italian cohort			
HapA	18.6	16.9[e]	0.937

Table 5 (*Continued*)

Study (author's name)	*Controls (%)*	*Cases (%)*	*P-values*
HapB	5.5	7.5[e]	*0.014*
SG13S25 *allele G*	92.6	92.0[e]	*0.682*
SG13S377 *allele G*	87.2	86.2[e]	*0.530*
SG13S114 *allele T*	63.3	63.0[e]	*0.921*
SG13S89 *allele G*	93.1	93.6[e]	*0.727*
SG13S32 *allele A*	48.7	47.6[e]	*0.620*
SG13S41 *allele A*	90.5	90.6[e]	*0.995*
SG13S35 *allele G*	91.7	90.5[e]	*0.396*

NS, nonsignificant.
[a]Myocardial infraction.
[b]Stroke.
[c]Control group for myocardial infraction.
[d]Control group for stroke.
[e]Coronary artery disease.
Italic numbers indicate the characteristics of case and control groups.

activation of monocyte/macrophages, as well as lipoxygenases-mediated LDL oxidation.

Peculiarities of the present study: strengths and limitations

Previous studies on *ALOX5AP* focused on MI patients *versus* controls selected from the general population or from event-free subjects such as in the prospective Physician's Health Study cohort.[4,13] MI is usually a late thrombotic complication superimposed on coronary atherosclerotic plaque rupture,[26] so that design of previous studies did not directly allow to separate a putative specific role of *ALOX5AP* in MI rather than in CAD development. Our alternative experimental design focused on subjects with angiographically proven CAD, with or without a previous documented MI. Moreover, the angiography-based design enabled us to select CAD-free subjects with an objectively defined control status, a critical issue in genetic association studies.[27] This allowed us to overcome the caveat, common in Western general populations where atherosclerosis is endemic, of enrolling controls with substantial coronary atherosclerotic lesions, although not yet clinically evident. While our CAD-free subjects cannot be considered a 'typical' control group, we feel confident about their acceptable representativity of the background general population, being their genotype and haplotype distributions, not fundamentally different from those observed in controls from German and US populations (see above). Since we noted haplotype differences only between the whole CAD group *versus* the CAD-free group, and not between CAD patients with or without MI, our data appear to be consistent with a more relevant role of *ALOX5AP* in atherogenesis rather than in thrombogenesis, according to many of the above-mentioned biochemical data.

This study suffers of common limitations of genetic association studies with complex traits.[28] Despite the unbalance between case and controls, it was sufficiently powered to detect a predefined effect of *ALOX5AP* haplotypes on CAD (see above). On the other hand, we could not properly analyse some interesting issues such as a possible stronger effect of *ALOX5AP* in males than in females,[4] because of the limited number of women enrolled.

Finally, from a possible practical perspective it has to be taken into account the relatively poor frequency of 'at-risk' haplotypes in our population, as well as their modest effect on the CAD risk.

Conclusions

ALOX5AP represents the paradigm of a new class of promising genes identified by powerful genome-wide investigations, which is currently an object of intense investigations to confirm their role in CAD susceptibility. Our data neither refute nor strongly support this hypothesis. Adding them to current knowledge, some evidence on *ALOX5AP* as a genetic susceptibility factor for CAD has now emerged in four out of five independent populations (Icelandic, British, Japanese, and Italian; but not in North America). Our angiography-based study suggests a possible role of *ALOX5AP*/FLAP in the development of the atheroma rather than in its late thrombotic complications such as MI. Such a role, if any, appears to be modest. Much further work is needed to understand the reason(s) for heterogeneous results, as well as to identify possible *ALOX5AP* pathogenic variations.

Acknowledgements

This work was supported by grants from the Veneto Region, Italian Ministry of University and Research (Grant no. 2005/065152), and the Cariverona Foundation, Verona, Italy. We wish to thank Mrs Maria Zoppi for invaluable secretary help, and Mr Diego Minguzzi for technical assistance. The authors declare that they have no potential conflict of interests.

References

1 Watkins H, Farrall M: Genetic susceptibility to coronary artery disease: from promise to progress. *Nat Rev Genet* 2006; 7: 163–173.
2 Lusis AJ, Fogelman AM, Fonarow GC: Genetic basis of atherosclerosis, Part I, New genes and pathways. *Circulation* 2004; 110: 1868–1873.
3 Wang Q: Molecular genetics of coronary artery disease. *Curr Opin Cardiol* 2005; 20: 182–188.
4 Helgadottir A, Manolescu A, Thorleifsson G *et al*: The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nat Genet* 2004; 36: 233–239.
5 Dixon RAF, Diehl RE, Opas E *et al*: Requirement of a 5-lipoxygenase-activating protein for leukotriene synthesis. *Nature* 1990; 343: 282–284.
6 Miller DK, Gillard JW, Vickers PJ *et al*: Identification and isolation of a membrane protein necessary for leukotriene production. *Nature* 1990; 343: 278–281.

7 Spanbroek R, Grabner R, Lotzer K *et al*: Expanding expression of 5-lipoxygenase pathway within the arterial wall during human atherogenesis. *Proc Natl Acad Sci USA* 2003; **100**: 1238–1243.

8 Qiu H, Gabrielsen A, Agardh HE *et al*: Expression of 5-lipoxygenase and leukotriene A4 hydrolase in human atherosclerotic lesions correlates with symptoms of plaque instability. *Proc Natl Acad Sci USA* 2006; **103**: 8161–8166.

9 Mehrabian M, Allayee H, Wong J *et al*: Identification of 5-lipoxygenase as a major gene contributing to atherosclerosis susceptibility in mice. *Circ Res* 2002; **91**: 120–126.

10 Zhao L, Funk CD: Lipoxygenase pathways in atherogenesis. *Trends Cardiovasc Med* 2004; **14**: 191–195.

11 Löhmussaar E, Gschwendtner A, Mueller JC *et al*: ALOX5AP gene and the PDE4D gene in a central European population of stroke patients. *Stroke* 2005; **36**: 731–736.

12 Meschia JF, Brott TG, Brown RD *et al*: Phosphodiesterase 4D and 5-lipoxygenase activating protein in ischemic stroke. *Ann Neurol* 2005; **58**: 351–361.

13 Zee RY, Cheng S, Hegener HH, Erlich HA, Ridker PM: Genetic variants of arachidonate 5-lipoxygenase activating protein and risk of incident myocardial infarction and ischemic stroke. *Stroke* 2006; **37**: 2007–2011.

14 Girelli D, Russo C, Ferraresi P *et al*: Polymorphisms in the factor VII gene and the risk of myocardial infarction in patients with coronary artery disease. *N Engl J Med* 2000; **343**: 774–780.

15 Devlin B, Risch N: A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 1995; **29**: 311–322.

16 Lake SL, Lyon H, Tantisira K *et al*: Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered* 2003; **55**: 56–65.

17 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.

18 Kajimoto K, Shioji K, Ishida C *et al*: Validation of the association between the gene encoding 5-lipoxygenase-activating protein and myocardial infarction in a Japanese population. *Circ J* 2005; **69**: 1029–1034.

19 Hansson GK: Inflammation, atherosclerosis, and coronary artery disease. *N Engl J Med* 2005; **352**: 1685–1695.

20 Samuelsson B: Leukotrienes: mediators of immediate hypersensitivity reactions and inflammation. *Science* 1983; **220**: 568–575.

21 Helgadottir A, Manolescu A, Helgason A *et al*: A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nat Genet* 2006; **38**: 68–74.

22 Dwyer JH, Allayee H, Dwyer KM *et al*: Arachidonate 5-lipoxigenase promoter genotype, dietary arachidonic acid and atherosclerosis. *N Engl J Med* 2004; **350**: 29–37.

23 Funk CD: Leukotriene modifiers as potential therapeutics for cardiovascular disease. *Nat Rev Drug Discov* 2005; **4**: 664–672.

24 Steinberg D: At last, direct evidence that lipoxygenases play a role in atherogenesis. *J Clin Invest* 1999; **103**: 1487–1488.

25 Lötzer K, Funk CD, Habenicht AJR: The 5-lipoxygenase pathway in arterial wall biology and atherosclerosis. *Biochim Biophys Acta* 2005; **1736**: 30–37.

26 Lusis AJ: Atherosclerosis. *Nature* 2000; **407**: 233–241.

27 Lander ES, Schork NJ: Genetic dissection of complex traits. *Science* 1994; **265**: 2037–2048.

28 Colhoun HM, McKeigue PM, Davey Smith G: Problems of reporting genetic associations with complex outcomes. *Lancet* 2003; **361**: 865–872.

29 Helgadottir A, Gretarsdottir S, St Clair D *et al*: Association between the gene encoding 5-lipoxygenase-activating protein and stroke replicated in a Scottish population. *Am J Hum Genet* 2005; **76**: 505–509.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

Genetic Variants of Arachidonate 5-Lipoxygenase–Activating Protein, and Risk of Incident Myocardial Infarction and Ischemic Stroke

A Nested Case-Control Approach

Robert Y.L. Zee, PhD; Suzanne Cheng, PhD; Hillary H Hegener, BS; Henry A. Erlich, PhD; Paul M Ridker, MD

Background and Purpose—Recent findings have implicated specific gene polymorphisms of arachidonate 5-lipoxygenase–activating protein (ALOX5AP), and 2 at-risk haplotypes (HapA, HapB) in myocardial infarction and stroke. To date, no prospective data are available.

Methods—We evaluated 10 specific Icelandic ALOX5AP gene variants among 600 male participants with incident atherothrombotic events (myocardial infarction [MI] or ischemic stroke) and among 600 age- and smoking-matched male participants, all white, who remained free of reported cardiovascular disease during follow-up within the Physicians' Health Study cohort.

Results—Overall allele, genotype, and haplotype distributions were similar between cases and controls. Single-marker conditional logistic regression analysis adjusted for potential risk factors found no association with risk of atherothrombotic events. Further investigation using a haplotype-based approach showed similar null findings with MI (HapA: odds ratio [OR]=1.18, 95% CI, 0.76 to 1.85; P=0.46; HapB: odds ratio=0.62, 95% CI, 0.36 to 1.07; P=0.08), and with ischemic stroke (HapA: odds ratio=1.11, 95% CI, 0.65 to 1.89; P=0.71; HapB: odds ratio=0.82, 95% CI, 0.47 to 1.42; P=0.47).

Conclusions—We found no evidence for an association of the specific Icelandic ALOX5P gene variants/at-risk haplotypes tested with risk of incident MI nor ischemic stroke in this prospective, non-Icelandic study. (***Stroke*. 2006;37:2007-2011.**)

Key Words: ALOX5AP ■ haplotypes ■ MI ■ risk factors ■ stroke

Cardiovascular diseases, including myocardial infarction (MI) and ischemic stroke, are the leading causes of mortality and morbidity in western countries. The underlying pathogenesis is likely to be mediated by both genetic and environmental risk factors. The initial report,[1] in an Icelandic population, of a significant association of genetic variants of arachidonate 5-lipoxygenase–activating protein (ALOX5AP) with increased risk of MI and stroke has attracted great interest. In their study, Helgadottir and coauthors reported a linkage and association of a 4-single-nucleotide polymorphism (SNP) haplotype, HapA, of *ALOX5AP* gene with risk of MI and stroke.[1] In addition, they reported an association of a different 4-SNP haplotype, HapB, with risk of MI in a British population.[1] Helgadottir and coauthors further assessed the contribution of *ALOX5AP* variants, in particular the HapA, and HapB haplotypes, to stroke, in a Scottish population, and found that the HapA haplotype confers a relative risk of 1.36 assuming a multiplicative model (P=0.007) for stroke.[2] However, they found no association for HapB. Subsequent studies by others in several non-Icelandic populations have since yielded conflicting results.[3,4]

To date, no prospective genetic-epidemiological data are available on risk of MI, and ischemic stroke. We therefore simultaneously evaluated the role of 10 ALOX5AP (GeneID: 241; Chromosome: 13q12) SNPs (SG13S25, SG13S377, SG13S106, SG13S114, SG13S89, SG13S30, SG13S32, SG13S41, SG13S42, and SG13S35), and specific haplotypes thereof, in particular HapA, and HapB at-risk haplotypes, as risk determinants of incident MI, and ischemic stroke in a prospective, nested case-control sample within the Physicians' Health Study (PHS) cohort. These polymorphisms (except SG13S106, SG13S30, and SG13S42: unpublished data from deCODE Genetics) were chosen based on the associations observed in the Icelandic study.[1]

Materials and Methods

Study Design

We used a nested case-control design within the PHS,[5] a randomized, double-blinded, placebo-controlled trial of aspirin and beta carotene initiated in 1982 among 22 071 males, predominantly white

Received March 31, 2006; accepted May 2, 2006.

From the Center for Cardiovascular Disease Prevention, the Donald W. Reynolds Center for Cardiovascular Research, the Leducq Center for Molecular and Genetic Epidemiology (R.Y.L.Z., H.H.H., P.M.R.), Brigham and Women's Hospital, Harvard Medical School, Boston, Mass; and the Roche Molecular Systems (S.C., H.A.E.), Alameda, Calif.

Correspondence to Dr Robert Y.L. Zee, Laboratory of Genetic and Molecular Epidemiology, Center for Cardiovascular Disease Prevention, Brigham and Women's Hospital, 900 Commonwealth Ave East, Boston, MA 02215. E-mail rzee@rics.bwh.harvard.edu

Stroke is available at http://www.strokeaha.org **DOI: 10.1161/01.STR.0000229905.25080.01**

(>94%), US physicians, 40 to 84 years of age at study entry. Before randomization, 14 916 participants provided an EDTA-anticoagulated blood sample and stored for genetic analysis. All participants were free of prior MI, stroke, transient ischemic attacks, and cancer at study entry. As the study participants were all US male physicians, yearly follow-up self-report questionnaires provide reliable updated information on newly developed diseases and the presence or absence of other cardiovascular risk factors. History of cardiovascular risk factors, such as hypertension (>140/90 mm Hg or on antihypertensive medication), diabetes or hyperlipidemia (>240 mg/dL), was defined by self-report of diagnosis at entry into the study. For all reported incident vascular events occurring after study enrollment, hospital records, death certificates, and autopsy reports were requested and reviewed by an end-points committee using standardized diagnostic criteria.

The diagnosis of MI was confirmed by evidence of symptoms in the presence of either diagnostic elevations of cardiac enzymes or diagnostic changes on electrocardiograms. In the case of fatal events, the diagnosis of MI was also accepted based on autopsy findings. Stroke was defined by the presence of a new focal neurological deficit, with symptoms and signs persisting for >24 hours, and was ascertained from blinded review of medical records, autopsy results and the judgment of a board-certified neurologist, on the basis of clinical reports, computed tomographic, or MRI scanning.

For each case (MI or ischemic stroke), a control matched by age, smoking history (never, past, or current) and length of follow-up were chosen among those subjects who remained free of vascular diseases. The present association study consisted of 341 MI case-control pairs, and 259 ischemic stroke case-control pairs, all white males.

The study was approved by the Brigham and Women's Hospital Institutional Review Board for Human Subjects Research.

Genotyping Determination

Genotyping was performed using an immobilized probe approach, as previously described (Roche Molecular Systems).[6] In brief, each DNA sample was amplified in a multiplex polymerase chain reaction using biotinylated primers. Each polymerase chain reaction product pool was then hybridized to a panel of sequence-specific oligonucleotide probes immobilized in a linear array. The colorimetric detection method was based on the use of streptavidin-horseradish peroxidase conjugate with hydrogen peroxide and 3,3′,5,5′-tetramethylbenzidine as substrates.

To confirm genotype assignment, scoring was carried out by 2 independent observers. Discordant results (<1% of all scoring) were resolved by a joint reading, and where necessary, a repeat genotyping. Results were scored blinded as to case-control status. Overall completion rate of genotyping determination was ≥95%.

Statistical Analysis

Allele and genotype frequencies among cases and controls were compared with values predicted by Hardy-Weinberg equilibrium using the χ^2 test. Relative risks associated with each genotype were calculated separately by conditional logistic regression analysis conditioning on the matching by age, smoking status, and length of follow-up since randomization, and further controlling for randomized treatment assignment, history of hypertension, presence or absence of diabetes, and body mass index, assuming an additive, dominant, or recessive mode of inheritance. Pairwise linkage disequilibrium (LD) was examined as described by Devlin and Risch.[7] For comparison with published reports by others, we examined 2 previously described at-risk haplotypes: HapA (SG13S25*G*-SG13S114*T*-SG13S89*G*-SG13S32*A*), and HapB (SG13S377*A*-SG13S114*A*-SG13S41*A*-SG13S35*G*). Haplotype estimation and inference was determined using PHASE v2.1.[8,9] Haplotype distributions between cases and controls were examined by likelihood ratio test. The relationship between haplotypes and clinical outcomes was examined using a haplotype-based logistic regression analysis with baseline-parameterization,[10] adjusting for the same risk factors. All analyses were carried out using SAS/Genetics 9.1 package (SAS Institute, Inc). For each odds ratio (OR), we calculated 95% CIs. A 2-tailed *P* value of 0.05 was considered a statistically significant result.

TABLE 1. Baseline Characteristics of Study Participants Who Subsequently Developed Any Arterial Event (Cases), and Those Who Remained Free of Vascular Disease During Follow-Up (Controls)

	Controls (n=600)	Cases (n=600)	*P*
Age, y	60.8±0.3	61.0±0.3	m.v.
Smoking status, %			m.v.
Never	41.7	41.7	
Past	41.5	41.5	
Current	16.8	16.8	
Body mass index, kg/m^2	24.9±0.1	25.4±0.1	0.002
Blood pressure, mm Hg			
Systolic	128.6±0.5	132.7±0.6	<0.0001
Diastolic	79.6±0.3	81.8±0.3	<0.0001
Hyperlipidemia, %	14.9	22.8	<0.001
Hypertension, %	29.0	47.2	<0.0001
Diabetes, %	2.8	8.9	<0.0001
Aspirin use, %	46.3	44.8	0.61
Family history of premature CAD <60 years of age, %	8.9	10.9	0.24

Mean±SE unless otherwise stated.

m.v. indicates matching variable; CAD, coronary artery disease.

Continuous and categorical variables were tested by paired *t* test and McNemar test, respectively.

Results

Baseline characteristics of cases and controls are shown in Table 1. As expected, the case participants had a higher prevalence of traditional cardiovascular risk factors at baseline as compared with controls. The genotype frequencies for the polymorphisms tested were in Hardy-Weinberg equilibrium in the control group and in the case group.

Using a single-marker χ^2 analysis, allele and genotype distributions were similar between cases and controls (Table 2). Results from the adjusted conditional logistic regression analysis, assuming additive, dominant, or recessive mode of inheritance, showed no significant association of the variants tested with the clinical outcomes ($P \geq 0.07$; data not shown). In general, the polymorphisms tested were in LD (supplemental Table I, available online at http://stroke.ahajournals.org). The overall haplotype distributions between cases and controls were similar (MI: HapA region, $P=0.79$, HapB region, $P=0.94$; ischemic stroke: HapA region, $P=0.77$, HapB region, $P=0.26$; supplemental Table II, available online at http://stroke.ahajournals.org). The most frequent haplotypes were G-T-G-C, and G-T-A-G for HapA region, and HapB region, respectively (supplemental Table II), and thus were used as the referents. Results from the adjusted haplotype-based conditional logistic regression analysis again showed similar null findings (supplemental Table III, available online at http://stroke.ahajournals.org).

TABLE 2. Genotype and Allele Distribution

ALOX5AP Genotype, %	MI Controls	MI Cases	*P*	IsST Controls	IsST Cases	*P*
SG13S25			0.80			0.29
GG	81.31	80.56		83.13	79.58	
GA	18.07	19.14		15.64	20.00	
AA	0.62	0.31		1.23	0.42	
Allele			0.89			0.47
G	0.90	0.90		0.91	0.90	
A	0.10	0.10		0.09	0.10	
SG13S377			0.71			0.35
GG	75.39	78.09		70.37	75.42	
GA	23.05	20.68		25.93	22.50	
AA	1.56	1.23		3.70	2.08	
Allele			0.41			0.15
G	0.87	0.88		0.83	0.87	
A	0.13	0.12		0.17	0.13	
SG13S106			0.54			0.20
GG	50.16	46.60		45.27	45.00	
GA	37.69	41.98		44.86	40.00	
AA	12.15	11.42		9.88	15.00	
Allele			0.59			0.38
G	0.69	0.68		0.68	0.65	
A	0.31	0.32		0.32	0.35	
SG13S114			0.90			0.96
TT	47.04	45.37		41.56	42.08	
TA	41.43	42.28		43.62	42.50	
AA	11.53	12.35		14.81	15.42	
Allele			0.63			0.99
T	0.68	0.68		0.63	0.63	
A	0.32	0.32		0.37	0.37	
SG13S89			0.76			0.80
GG	89.72	88.89		89.71	89.17	
GA	9.66	10.80		9.47	10.42	
AA	0.62	0.31		0.82	0.42	
Allele			0.84			0.96
G	0.95	0.94		0.94	0.94	
A	0.05	0.06		0.06	0.06	
SG13S30			0.83			0.38
GG	58.57	58.95		51.85	57.92	
GT	37.69	36.42		41.15	36.67	
TT	3.74	4.63		7.00	5.42	
Allele			0.91			0.17
G	0.77	0.77		0.72	0.76	
T	0.23	0.23		0.28	0.24	
SG13S32			0.30			0.32
CC	27.73	22.84		24.28	20.83	
CA	52.96	54.63		47.33	54.17	
AA	19.31	22.53		28.40	25.00	
Allele			0.15			0.99
C	0.54	0.50		0.48	0.48	
A	0.46	0.50		0.52	0.52	

(Continued)

TABLE 2. Continued

ALOX5AP Genotype, %	MI Controls	MI Cases	*P*	IsST Controls	IsST Cases	*P*
SG13S41			0.50			0.89
AA	82.87	83.02		84.36	85.42	
AG	15.58	16.36		14.40	13.75	
GG	1.56	0.62		1.23	0.83	
Allele			0.73			0.68
A	0.91	0.91		0.92	0.92	
G	0.09	0.09		0.08	0.08	
SG13S42			0.17			0.36
AA	28.04	34.88		38.68	35.00	
AG	50.78	45.99		43.62	50.00	
GG	21.18	19.14		17.70	15.00	
Allele			0.11			0.88
A	0.53	0.58		0.60	0.60	
G	0.47	0.42		0.40	0.40	
SG13S35			0.08			0.50
GG	81.31	85.80		79.42	83.33	
GA	18.69	13.58		19.75	16.25	
AA	...	0.62		0.82	0.42	
Allele			0.21			0.26
G	0.91	0.93		0.89	0.91	
A	0.09	0.07		0.11	0.09	

IsST indicates ischemic stroke.
P value for χ^2 test.

Discussion

The present prospective investigation provides no evidence for an association of the specific gene variants, nor at-risk haplotypes of the *ALOX5AP* gene, previously suggested as genetic risk determinants, with MI or stroke in a non-Icelandic white population.

In the initial Icelandic report,[1] a 4-SNP haplotype (HapA) was found to be associated with a 2× greater risk of MI, and an almost 2× greater risk of stroke. The same group also reported an association of a different 4-SNP ALOX5AP haplotype (HapB) with risk of MI in a British sample population[1] (Table 3). A subsequent report by Helgadottir and coauthors found an association between HapA and an increased risk of ischemic stroke (relative risk=1.35; P=0.02), and an over-representation of HapB (relative risk=1.65; P=0.02) with ischemic stroke in a Scottish male sample population[2] (Table 3). Recently, Lohmussar and coauthors[3] reported that sequence variants in the *ALOX5AP* gene are significantly associated with stroke, particularly in males, in a Central European sample population. A nominally significant association with stroke was observed for SG13S114 (OR=1.24; P=0.017), and SG13S100 (OR=1.26; P=0.024). However, they found no association of HapA with stroke risk.[3] More recently, Meschia and coauthors conducted the first replication study using a North American sample

TABLE 3. Summary of ALOX5AP At-Risk-Haplotypes Association Studies

	HapA		HapB	
	MI	Stroke	MI	Stroke
	Conf, Casf, R, *P*	Conf, Casf, R, *P*	Conf, Casf, R, *P*	Conf, Casf, R, *P*
Present study United States	0.14, 0.17, 1.18, 0.46	0.18, 0.15, 1.11, 0.71	0.07, 0.06, 0.62, 0.08	0.08, 0.07, 0.82, 0.47
Iceland[1]	0.10, 0.16, 1.80, <0.0001	0.10, 0.15, 1.67, <0.0001	Not available	*0.07, 0.07, 1.09, ns
United Kingdom[1]	0.15, 0.17, ns	Not available	0.04, 0.08, 1.95, 0.00037	Not available
Scotland[2]	Not available	0.14, 0.18, 1.35, 0.02	Not available	0.06, 0.09, 1.65, 0.02
Germany[3]	Not available	0.15, 0.15, ns	Not available	ns (data not shown)
North America[4]	Not available	ns (data not shown)	Not available	Not available

Conf indicates haplotype frequency in controls; Casf, haplotype frequency in cases; R, risk estimate; ns, nonsignificance.
HapA=SG13S25*G*-SG13S114*T*-SG13S89*G*-SG13S32*A*. HapB=SG13S377*A*-SG13S114*A*-SG13S41*A*-SG13S35*G*.
*Data extracted from reference 2.

population, and found no association between ALOX5AP gene variants and stroke, although MI was not investigated in their study.

Given this situation, a possible explanation for the apparent discrepancies is that the observed allele, genotype, and at-risk haplotype frequencies for the SNPs examined may differ between studies, which could be the result of population/ethnic differences. As previously suggested,[3,4] the *ALOX5AP* gene variation may play a substantial role in risk of MI, and stroke in Iceland (an isolate population), but a lesser role in non-Icelandic populations because of different population LD structures. These recent results are consistent with the initial report that different at-risk haplotypes were found between the Icelandic and British study populations.[1]

As shown in Table 3, not all of the published reports examined the same set of SNPs, nor did all of the reported studies examine the association of *ALOX5AP* variants with MI and stroke simultaneously. Further, not all published studies presented information on allele, genotype and at-risk haplotype frequencies, LD structure, and risk estimates, thus making a direct comparison and informative interpretation across studies difficult.

It has been noted in the initial report[1] that variants of *ALOX5AP* gene are involved in the pathophysiology of MI and stroke by increasing the production of leukotriene B4, a critical regulator in the 5-lipoxygenase pathway, and a proinflammatory agent. Leukotrienes are arachidonic acid metabolites, which have been implicated in various inflammatory conditions, including asthma, arthritis, psoriasis, and atherosclerosis.[11,12] Notably, a recent article by the same Icelandic group found a haplotype (HapK) of the gene encoding leukotriene A4 hydrolase, a protein in the same biochemical pathway of ALOX5AP, confers ethnicity-specific (particularly in blacks) risk of MI.[13]

The prospective nature of the PHS study and the use of a closed population sampling scheme in which subsequent case status was determined solely by the development of disease strongly reduce the possibility that our findings are attributable to bias or confounding. Our study cohort consists of entirely white males with distinct socioeconomic status (physicians), so our data cannot be generalized to other ethnic groups and women. In our study, we had the ability to detect, based on the present sample sizes, assuming 80% power, at an α of 0.05, a risk ratio of >1.54 (MI), and 1.64 (ischemic stroke) if the minor allele frequency is 0.50, and of >2.26 (MI), and 2.49 (ischemic stroke) if the minor allele frequency is 0.05 assuming a univariable-additive mode. Thus, we cannot rule out a modest risk of cardiovascular disease associated with the polymorphisms/haplotypes tested. It is important to recognize that association studies like this one can only examine the possible association between phenotype and the tested polymorphisms. Our study therefore cannot exclude the possibility that examination of different polymorphisms/loci, which would by definition have to be in linkage disequilibrium with the ones tested, might obtain different results.

In conclusion, our prospective study found no evidence for an association of specific Icelandic *ALOX5AP* gene polymorphisms/at-risk haplotypes examined with risk of atherothrombotic events. If corroborated in other non-Icelandic prospective studies, our data suggest that *ALOX5AP* gene variation is not informative for risk assessment of atherothrombosis in non-Icelandic populations.

Acknowledgments

We thank Michael Grow and Houman Khakpour (both at Roche Molecular Systems) for their expertise and efforts in developing the genotyping reagents used for this study. We thank Anna Helgadottir at deCODE Genetics for sharing unpublished data.

Sources of Funding

This work was supported by grants from the National Heart Lung and Blood Institute (HL-58755 and HL-63293), the Doris Duke Charitable Foundation, the American Heart Association, and the Donald W. Reynolds Foundation, Las Vegas, Nevada.

Disclosures

None.

References

1. Helgadottir A, Manolescu A, Thorleifsson G, Gretarsdottir S, Jonsdottir H, Thorsteinsdottir U, Samani NJ, Gudmundsson G, Grant SF, Thorgeirsson G, Sveinbjornsdottir S, Valdimarsson EM, Matthiasson SE, Johannsson H, Gudmundsdottir O, Gurney ME, Sainz J, Thorhallsdottir M, Andresdottir M, Frigge ML, Topol EJ, Kong A, Gudnason V, Hakonarson H, Gulcher JR, Stefansson K. The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nat Genet.* 2004;36:233–239.
2. Helgadottir A, Gretarsdottir S, St Clair D, Manolescu A, Cheung J, Thorleifsson G, Pasdar A, Grant SF, Whalley LJ, Hakonarson H, Thorsteinsdottir U, Kong A, Gulcher J, Stefansson K, MacLeod MJ. Association between the gene encoding 5-lipoxygenase-activating protein and stroke replicated in a scottish population. *Am J Hum Genet.* 2005;76:505–509.
3. Lohmussaar E, Gschwendtner A, Mueller JC, Org T, Wichmann E, Hamann G, Meitinger T, Dichgans M. *AlOX5AP* gene and the *PDE4D* gene in a central European population of stroke patients. *Stroke.* 2005; 36:731–736.
4. Meschia JF, Brott TG, Brown RD Jr, Crook R, Worrall BB, Kissela B, Brown WM, Rich SS, Case LD, Evans EW, Hague S, Singleton A, Hardy J. Phosphodiesterase 4d and 5-lipoxygenase activating protein in ischemic stroke. *Ann Neurol.* 2005;58:351–361.
5. Physician's health study. Aspirin and primary prevention of coronary heart disease. *N Engl J Med.* 1989;321:1825–1828.
6. Cheng S, Grow MA, Pallaud C, Klitz W, Erlich HA, Visvikis S, Chen JJ, Pullinger CR, Malloy MJ, Siest G, Kane JP. A multilocus genotyping assay for candidate markers of cardiovascular disease risk. *Genome Res.* 1999;9:936–949.
7. Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics.* 1995;29:311–322.
8. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 2001;68:978–989.
9. Stephens M, Donnelly P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet.* 2003;73:1162–1169.
10. Wallenstein S, Hodge SE, Weston A. Logistic regression model for analyzing extended haplotype data. *Genet Epidemiol.* 1998;15:173–181.
11. Samuelsson B. Leukotrienes: mediators of immediate hypersensitivity reactions and inflammation. *Science.* 1983;220:568–575.
12. Spanbroek R, Grabner R, Lotzer K, Hildner M, Urbach A, Ruhling K, Moos MP, Kaiser B, Cohnert TU, Wahlers T, Zieske A, Plenz G, Robenek H, Salbach P, Kuhn H, Radmark O, Samuelsson B, Habenicht AJ. Expanding expression of the 5-lipoxygenase pathway within the arterial wall during human atherogenesis. *Proc Natl Acad Sci U S A.* 2003;100:1238–1243.
13. Helgadottir A, Manolescu A, Helgason A, Thorleifsson G, Thorsteinsdottir U, Gudbjartsson DF, Gretarsdottir S, Magnusson KP, Gudmundsson G, Hicks A, Jonsson T, Grant SF, Sainz J, O'Brien SJ, Sveinbjornsdottir S, Valdimarsson EM, Matthiasson SE, Levey AI, Abramson JL, Reilly MP, Vaccarino V, Wolfe ML, Gudnason V, Quyyumi AA, Topol EJ, Rader DJ, Thorgeirsson G, Gulcher JR, Hakonarson H, Kong A, Stefansson K. A variant of the gene encoding leukotriene a4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nat Genet.* 2006;38:68–74.

No association of polymorphisms in the gene encoding 5-lipoxygenase-activating protein and myocardial infarction in a large central European population

Werner Koch, PhD[1], Petra Hoppmann, MD[1], Jakob C. Mueller, PhD[2], Albert Schömig, MD[1], and Adnan Kastrati, MD[1]

Purpose: Haplotypes based on polymorphisms in the gene encoding 5-lipoxygenase-activating protein have been linked with susceptibility to myocardial infarction in Iceland and the United Kingdom. We sought to replicate these association findings in a large case-control sample from Germany. **Methods:** The case group included 3657 patients with myocardial infarction and the control group comprised 1211 individuals with angiographically normal coronary arteries and without clinical signs or symptoms of myocardial infarction. Nine different polymorphisms were genotyped with the use of the TaqMan technique. **Results:** Genotype, allele, and haplotype analyses did not reveal significant associations between the polymorphisms and myocardial infarction. The negative results included a four-marker haplotype, termed HapA haplotype (odds ratio = 1.10; 95% confidence interval: 0.96–1.25), that was previously found to be related with myocardial infarction in a sample from Iceland, and a different four-marker haplotype, termed HapB haplotype (odds ratio = 0.94; 95% CI: 0.79–1.12), that was previously linked with myocardial infarction in a sample from the United Kingdom. Nine-marker haplotypes were not significantly associated with myocardial infarction in multiple logistic regression models adjusted for covariates ($P \geq 0.38$). **Conclusion:** In this sample from central Europe, specific polymorphisms in the gene for 5-lipoxygenase–activating protein were not associated with myocardial infarction, a result contrasting previous positive findings. ***Genet Med*** **2007:9(2):123–129.**

Key Words: *ALOX5AP, 5-lipoxygenase–activating protein (FLAP), genetics, haplotype, myocardial infarction*

Specific allelic forms of the gene encoding 5-lipoxygenase-activating protein (FLAP) have been linked with susceptibility to myocardial infarction (MI) and stroke.[1–4] These association findings may reflect a possible relationship of the regulatory function of FLAP in the inflammatory 5-lipoxygenase pathway and the important role attributed to inflammatory processes in atherosclerotic diseases.[5–9] The 5-lipoxygenase cascade leads to the formation of leukotrienes, which exhibit strong proinflammatory activities in cardiovascular tissues.[9–11] This pathway is especially active in arterial walls of patients afflicted with various lesion stages of atherosclerosis of the aorta and of coronary and carotid arteries.[10]

The gene for FLAP (*ALOX5AP*) contains five exons and spans approximately 31 kb in the chromosome 13q12 region.[12,13] Specific single nucleotide polymorphisms (SNPs), named SG13S100, SG13S106, SG13S114, and a four-marker haplotype of *ALOX5AP*, termed HapA haplotype (SG13S25-G, SG13S114-T, SG13S89-G, SG13S32-A), were found to be related to MI in a population sample from Iceland.[1] A different four-marker haplotype of *ALOX5AP*, termed HapB haplotype (SG13S377-A, SG13S114-A, SG13S41-A, SG13S35-G), but not the HapA haplotype, was linked with MI in a sample from the United Kingdom (UK).[1] No evidence of an association of the HapA or HapB haplotype with MI was obtained in a sample of white male physicians from the United States (US).[14]

We examined whether the nine different SNPs mentioned above, nine-marker haplotypes of these SNPs, and the HapA and HapB haplotypes were associated with MI in a German population. The sample consisted of 3657 patients with MI and 1211 control individuals, all of whom were assessed with coronary angiography.

METHODS

Patients and controls

Participants were recruited from Southern Germany and examined at Deutsches Herzzentrum München or 1. Medizinische

From the [1]Deutsches Herzzentrum München and 1. Medizinische Klinik, Klinikum rechts der Isar, Technische Universität München, Munich, Germany; [2]Max Planck Institute for Ornithology, Department of Behavioural Ecology and Evolutionary Genetics, Starnberg (Seewiesen), Germany.

The authors declare no conflict of interest.

Werner Koch, PhD, German Heart Center Munich, Lazarettstrasse 36, 80636 München, Germany. E-mail: wkoch@dhm.mhn.de

A supplementary Appendix is available via the ArticlePlus feature at www.geneticsinmedicine.org. Please go to the February issue and click on the ArticlePlus link posted with the article in the Table of Contents to view this material.

Submitted for publication September 27, 2006.

Accepted for publication December 5, 2006.

DOI: 10.1097/GIM.0b013e318030c9c5

Klinik rechts der Isar der Technischen Universität München from 1993 to 2002. After catheterization, 5264 individuals were deemed eligible for inclusion in the MI or control group. Written informed consent for genetic analysis was obtained from 97.1% (n = 5111) of these individuals. In no case was consent withdrawn. Blood samples assigned for DNA preparation had been collected from 95.2% (n = 4868) of the individuals who agreed to participate in the study. These individuals, 3657 patients with MI and 1211 controls, constituted the study population. Complete genotype data were obtained from all these patients and control individuals. The study protocol was approved by the Institutional Ethics committee and the reported investigations were in accordance with the principles of the Declaration of Helsinki.[15]

Definitions

Individuals were considered disease free and, therefore, eligible as controls when their coronary arteries were angiographically normal and when they had no history of MI, no symptoms suggestive of MI, no electrocardiographic signs of MI, and no regional wall motion abnormalities. Coronary angiography in the control individuals was performed for the evaluation of chest pain. The diagnosis of MI was established in the presence of chest pain lasting longer than 20 minutes combined with ST-segment elevation or pathologic Q waves on a surface electrocardiogram. Patients with MI had to show either an angiographically occluded infarct-related artery or regional wall motion abnormalities corresponding to the electrocardiographic infarct localization, or both. Systemic arterial hypertension was defined as a systolic blood pressure of ≥140 mm Hg and/or a diastolic blood pressure of ≥90 mm Hg,[16] on at least two separate occasions or antihypertensive treatment. Hypercholesterolemia was defined as a documented total cholesterol value ≥240 mg/dL (≥6.2 mmol/L) or current treatment with cholesterol-lowering medication. Persons reporting regular smoking in the previous 6 months were considered as current smokers. Diabetes mellitus was defined as the presence of an active treatment with insulin or an oral antidiabetic agent; for patients on dietary treatment, documentation of an abnormal fasting blood glucose or glucose tolerance test based on the World Health Organization criteria[17] was required for establishing this diagnosis.

Genetic analysis

Genomic DNA was extracted from peripheral blood leukocytes with the QIAamp DNA Blood Kit (Qiagen, Hilden, Germany) or the High Pure PCR Template Preparation Kit (Roche Applied Science, Mannheim, Germany). We designed and used TaqMan allelic discrimination assays for genotype analysis of nine SNPs in *ALOX5AP* (Table 1). Primers and probes (Table 1) were synthesized by Applied Biosystems (Darmstadt,

Table 1
SNPs in the *ALOX5AP* genomic region and nucleotide sequences of primers and probes used in TaqMan reactions

deCODE SNP ID[a]	NCBI dbSNP ID[b]	SNP bases	Position in AL512642[c]	Location	Primer (5′ → 3′)	Probe (5′ → 3′)[d]
SG13S25	—	G>A	26663	Upstream of exon 1	TCTGACAGCATCAGCTAGTCTCTTTC AAATTCATGTTGCTGTGTCCATACA	FAM-CACTGTTGCCCAGTGG VIC-AGCCACTGTTACCCAGT
SG13S377	—	G>A	31075	Upstream of exon 1	TTTGGCCAGACTGTCTTGAACTC TGGCTCATGCCTATAATCACAAAA	FAM-CCTGCCTCGGCCT VIC-CTGCCTCAGCCTC
SG13S100	rs4073259	A>G	33381	Upstream of exon 1	GGTGAAGTGGACTCCCTCCAT CCCCGCTCTGAGCTCCTT	FAM-AGCCAGCGCGCAG VIC-CAGCCAGTGCGCAG
SG13S106	rs9579646	G>A	37689	Intron 1	TGTGTAGAGCTGTCTTCCTAAAGTTCTG AAGCCACTGGAGATAGTTATGAAAGTG	FAM-AGTTAGGGCTGCCTC VIC-AGTTAGGACTGCCTCAG
SG13S114	rs10507391	T>A	39206	Intron 1	CCAGATGTATGTCCAAGCCTCTCT CTCTGTAAGGTAGGTCTATGGTTGCAA	FAM-TGCAATTCTAATTAACCTC VIC-TGCAATTCTATTTAACCTC
SG13S89	rs4769874	G>A	53551	Intron 3	TCGGGAGGCCGTGTTTC CCAGGGAGCAAGCATTAGCA	FAM-ATTATCACACGCGCTCT VIC-TATCACATGCGCTCTG
SG13S32	rs9551963	A>C	59657	Intron 4	CTGCTTTAGTTCTTGACCTCACCAA CTGGGGTTCAAGAGAGAAATTCC	FAM-AAGGATCTCATCTAGCAAT VIC-AAGGATCTCATCGAGCAA
SG13S41	rs9315050	A>G	63155	Intron 4	CCTGTCTCCAAATACAGTCCCATT AGGTCCCTTCCAAAATTCATATGTT	FAM-ATCTTTACTCTCAGTTCCT VIC-TCTTTACCCTCAGTTCC
SG13S35	—	G>A	67227	Downstream of exon 5	CCTGGCATTGAGGAGTTTTCC ACCCCACAAATACCTACAAATATGTGTAT	FAM-TAAAAAACCGAAAGGAC VIC-TTAAAAAACTGAAAGGACC

[a] Helgadottir et al.[1]
[b] NCBI SNP database (http://www.ncbi.nlm.nih.gov/entrez/); last accessed September 27, 2006.
[c] NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/entrez/); sequence version of May 18, 2005.
[d] FAM (6-carboxy-fluorescein) or VIC (proprietary dye of Applied Biosystems) was attached to the 5′ ends of the probe oligonucleotides. The sequences of the probes used for analysis of the SG13S25, SG13S377, SG13S106, and SG13S114 SNPs corresponded to the coding strand and the sequences of the probes used for analysis of the SG13S100, SG13S89, SG13S32, SG13S41, and SG13S35 SNPs corresponded to the noncoding strand; the allele-specific nucleotide in each probe sequence is underlined.
SNPs, single nucleotide polymorphisms.

Table 2

Baseline characteristics of the control group and the MI group

Characteristic	Control group (n = 1211)	MI group (n = 3657)	P
Age, yr	60.3 ± 11.9	64.0 ± 12.0	<0.0001
Women	598 (49.4)	885 (24.2)	<0.0001
Arterial hypertension	589 (48.6)	2246 (61.4)	<0.0001
Hypercholesterolemia	602 (49.7)	2067 (56.5)	<0.0001
Current cigarette smoking	184 (15.2)	1849 (50.6)	<0.0001
Diabetes mellitus	65 (5.4)	754 (20.6)	<0.0001

Age is mean ± SD; other variables are presented as number (%).
MI, myocardial infarction.

Germany). To accomplish allele-specific signaling, the probes contained the fluorogenic dyes 6-carboxy-fluorescein (FAM) or VIC (proprietary dye of Applied Biosystems). Minor groove binder groups were conjugated with the 3′ ends of the oligonucleotides to facilitate formation of stable duplexes between the probes and their single-stranded DNA targets.[18] Approximately 20% of the DNA samples were retyped with each TaqMan system to control for correct sample handling and data acquisition. The results of these repeat assays were in full agreement with the original genotyping results.

Analyses of PCR products with allele-discriminating restriction enzymes and/or DNA sequencing were used to verify the accuracy of TaqMan genotyping. We employed the restriction enzymes BglI (SG13S25 SNP), HaeIII (SG13S377 SNP), NsbI (SG13S100 SNP), SatI (SG13S106 SNP), TasI (SG13S114

Table 3

Genotype distributions and allele frequencies of *ALOX5AP* SNPs in the control group and the MI group

deCODE SNP ID	Genotype	Control group (1211 genotypes)	MI group (3657 genotypes)	P	Allele	Control group (2422 alleles)	MI group (7314 alleles)	P
SG13S25	GG	963 (79.5)	2949 (80.6)	0.54	G	2162 (89.3)	6579 (90.0)	0.33
	GA	236 (19.5)	681 (18.6)		A	260 (10.7)	735 (10.0)	
	AA	12 (1.0)	27 (0.7)					
SG13S377	GG	861 (71.1)	2714 (74.2)	0.053	G	2047 (84.5)	6285 (85.9)	0.086
	GA	325 (26.8)	857 (23.4)		A	375 (15.5)	1029 (14.1)	
	AA	25 (2.1)	86 (2.4)					
SG13S100	AA	494 (40.8)	1461 (40.0)	0.11	A	1521 (62.8)	4636 (63.4)	0.60
	AG	533 (44.0)	1714 (46.9)		G	901 (37.2)	2678 (36.6)	
	GG	184 (15.2)	482 (13.2)					
SG13S106	GG	568 (46.9)	1697 (46.4)	0.27	G	1644 (67.9)	4998 (68.3)	0.68
	GA	508 (41.9)	1604 (43.9)		A	778 (32.1)	2316 (31.7)	
	AA	135 (11.1)	356 (9.7)					
SG13S114	TT	526 (43.4)	1591 (43.5)	0.40	T	1586 (65.5)	4842 (66.2)	0.52
	TA	534 (44.1)	1660 (45.4)		A	836 (34.5)	2472 (33.8)	
	AA	151 (12.5)	406 (11.1)					
SG13S89	GG	1093 (90.3)	3332 (91.1)	0.60	G	2301 (95.0)	6983 (95.5)	0.34
	GA	115 (9.5)	319 (8.7)		A	121 (5.0)	331 (4.5)	
	AA	3 (0.2)	6 (0.2)					
SG13S32	AA	301 (24.9)	924 (25.3)	0.39	A	1224 (50.5)	3650 (49.9)	0.59
	AC	622 (51.4)	1802 (49.3)		C	1198 (49.5)	3664 (50.1)	
	CC	288 (23.8)	931 (25.5)					
SG13S41	AA	1047 (86.5)	3166 (86.6)	0.96	A	2253 (93.0)	6810 (93.1)	0.88
	AG	159 (13.1)	478 (13.1)		G	169 (7.0)	504 (6.9)	
	GG	5 (0.4)	13 (0.4)					
SG13S35	GG	977 (80.7)	3025 (82.7)	0.24	G	2172 (89.7)	6645 (90.9)	0.086
	GA	218 (18.0)	595 (16.3)		A	250 (10.3)	669 (9.1)	
	AA	16 (1.3)	37 (1.0)					

Variables are presented as number (%) of genotypes or alleles in control individuals and myocardial infarction patients.
SNPs, single nucleotide polymorphisms.

SNP), XceI (SG13S89 SNP), TaqI (SG13S32 SNP), and BslLI (SG13S41 SNP) (MBI Fermentas). DNA sequencing was used to test whether one or more additional polymorphisms were present in the probe-binding section of the amplicons, because they may interfere with TaqMan reactions and result in wrong genotype assignments. With each SNP, 100 DNA samples were examined by sequencing. The known SNPs were identified as the only sequence variabilities in the probe-binding regions. Thus the probability of genotyping errors due to possible further sequence variations was relatively low.

Clinicians responsible for diagnosis were not aware of the genetic data. All genetic analyses were blinded.

Statistical analysis

The analysis consisted of comparisons of genotype, allele, and haplotype frequencies between the control group and the group of patients with MI. Because stronger associations of the HapA haplotype with MI were observed in men compared to women in both the Iceland and UK studies,[1] we also conducted separate analyses of SNP genotype distributions and HapA and HapB haplotype frequencies in the groups of men and women. Discrete variables are expressed as counts (%) and compared using the χ^2 test. Continuous variables are expressed as mean ± SD and compared by means of the unpaired, two-sided *t* test. Haplotypes were reconstructed from genotype data with the use of the software

Table 4
Genotype distributions of *ALOX5AP* SNPs in the women and men of the control group and the MI group

		Women			Men		
deCODE SNP ID	Genotype	Control group (n = 598)	MI group (n = 885)	*P*	Control group (n = 613)	MI group (n = 2772)	*P*
SG13S25	GG	465 (77.8%)	716 (80.9%)	0.13	498 (81.2%)	2233 (80.6%)	0.87
	GA	127 (21.2%)	166 (18.8%)		109 (17.8%)	515 (18.6%)	
	AA	6 (1.0%)	3 (0.3%)		6 (1.0%)	24 (0.9%)	
SG13S377	GG	429 (71.7%)	659 (74.5%)	0.37	432 (70.5%)	2055 (74.1%)	0.13
	GA	157 (26.3%)	205 (23.2%)		168 (27.4%)	652 (23.5%)	
	AA	12 (2.0%)	21 (2.4%)		13 (2.1%)	65 (2.3%)	
SG13S100	AA	255 (42.6%)	372 (42.0%)	0.64	239 (39.0%)	1089 (39.3%)	0.10
	AG	261 (43.6%)	404 (45.6%)		272 (44.4%)	1310 (47.3%)	
	GG	82 (13.7%)	109 (12.3%)		102 (16.6%)	373 (13.5%)	
SG13S106	GG	291 (48.7%)	431 (48.7%)	0.66	277 (45.2%)	1266 (45.7%)	0.27
	GA	247 (41.3%)	377 (42.6%)		261 (42.6%)	1227 (44.3%)	
	AA	60 (10.0%)	77 (8.7%)		75 (12.2%)	279 (10.1%)	
SG13S114	TT	270 (45.2%)	403 (45.5%)	0.65	256 (41.8%)	1188 (42.9%)	0.53
	TA	256 (42.8%)	389 (44.0%)		278 (45.4%)	1271 (45.9%)	
	AA	72 (12.0%)	93 (10.5%)		79 (12.9%)	313 (11.3%)	
SG13S89	GG	545 (91.1%)	813 (91.9%)	0.84	548 (89.4%)	2519 (90.9%)	0.37
	GA	52 (8.7%)	70 (7.9%)		63 (10.3%)	249 (9.0%)	
	AA	1 (0.2%)	2 (0.2%)		2 (0.3%)	4 (0.1%)	
SG13S32	AA	147 (24.6%)	221 (25.0%)	0.51	154 (25.1%)	703 (25.4%)	0.89
	AC	315 (52.7%)	442 (49.9%)		307 (50.1%)	1360 (49.1%)	
	CC	136 (22.7%)	222 (25.1%)		152 (24.8%)	709 (25.6%)	
SG13S41	AA	524 (87.6%)	773 (87.3%)	0.99	523 (85.3%)	2393 (86.3%)	0.75
	AG	72 (12.0%)	109 (12.3%)		87 (14.2%)	369 (13.3%)	
	GG	2 (0.3%)	3 (0.3%)		3 (0.5%)	10 (0.4%)	
SG13S35	GG	472 (78.9%)	722 (81.6%)	0.19	505 (82.4%)	2303 (83.1%)	0.59
	GA	114 (19.1%)	154 (17.4%)		104 (17.0%)	441 (15.9%)	
	AA	12 (2.0%)	9 (1.0%)		4 (0.7%)	28 (1.0%)	

Variables are presented as number (%) of genotypes in control individuals and MI patients.
SNPs, single nucleotide polymorphisms; MI, myocardial infarction.

package PHASE.[19] We tested for the independent association effect of nine-marker haplotypes in multiple logistic regression models of MI that included as covariates age, gender, history of arterial hypertension, history of hypercholesterolemia, current cigarette smoking, and diabetes mellitus. Adjusted odds ratios and 95% Wald confidence intervals were calculated based on these models. Statistical significance was set at $P < 0.05$.

RESULTS

The main baseline characteristics of the control group (n = 1211) and the group of patients with MI (n = 3657) are shown in Table 2. Mean age of the MI patients was higher than that of the control group; the proportion of women was lower in the patient group than in the control group; and history of arterial hypertension and hypercholesterolemia, current cigarette smoking, and diabetes mellitus were more prevalent in the MI patient group than in the control group ($P < 0.0001$ for all comparisons; Table 2).

Genotype distributions and allele frequencies of the *ALOX5AP* SNPs were not significantly different between the control group and patient group (Table 3). Significant sex-related differences of the genotype distributions were not found (Table 4).

Figure 1 shows the linkage disequilibrium (LD) block structure defined by the nine genotyped SNPs. Strong LD was present across the *ALOX5AP* region (Fig. 1). Frequencies of the HapA (SG13S25-G, SG13S114-T, SG13S89-G, SG13S32-A) and HapB (SG13S377-A, SG13S114-A, SG13S41-A, SG13S35-G) haplotypes were not substantially different between the control group and the patient group (Table 5). Risk estimates were 1.10 (95% CI: 0.96–1.25) for the HapA haplotype and 0.94 (95% CI: 0.79–1.12) for the HapB haplotype. Haplotypes defined by nine SNPs were not present at significantly different proportions among the control individuals and patients, with the exception of the Hap5 haplotype, which showed a moderately higher frequency in the control group than in the patient group (Table 6).

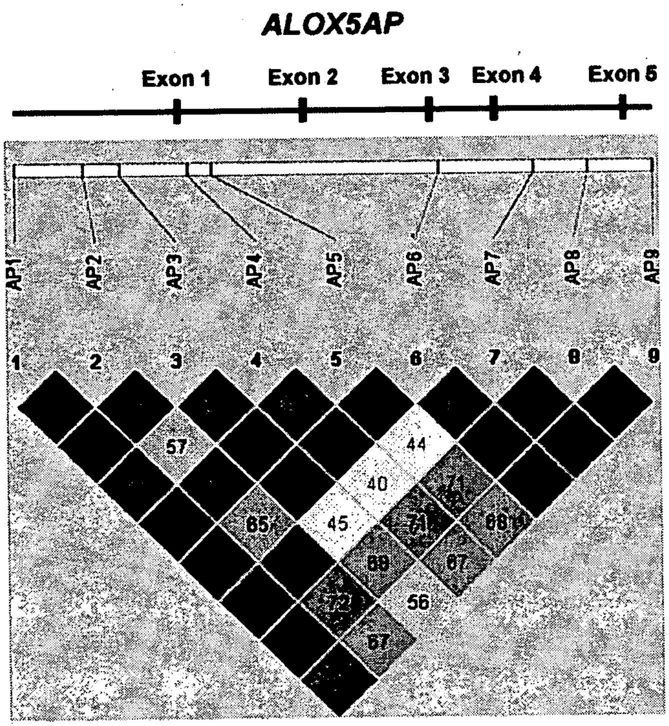


Fig. 1. Genetic diversity at the *ALOX5AP* genomic region located in the long arm of chromosome 13 (band q12). The exon-intron structure was adapted from sequence data deposited in the NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/entrez/) under accession number AL512642, version of May 18, 2005. The values within squares show the pairwise correlations between single nucleotide polymorphisms (SNPs) (measured as D') defined at the top left and top right sides of the squares. Squares without a number indicate $D' = 1.00$. SNP designations: AP1 = SG13S25, AP2 = SG13S377, AP3 = SG13S100, AP4 = SG13S106, AP5 = SG13S114, AP6 = SG13S89, AP7 = SG13S32, AP8 = SG13S41, AP9 = SG13S35.

Table 5

Frequencies of the HapA and HapB haplotypes in the control and MI groups

Haplotype	Control group (2422 haplotypes)	MI group (7314 haplotypes)	*P*
HapA	359 (14.8)	1171 (16.0)	0.16
HapB	182 (7.5)	518 (7.1)	0.48

Haplotype frequencies are presented as number (%). The HapA haplotype is defined by the alleles SG13S25-G, SG13S114-T, SG13S89-G, and SG13S32-A, and the HapB haplotype is defined by the alleles SG13S377-A, SG13S114-A, SG13S41-A, and SG13S35-G (Helgadottir et al.[1]).
MI, myocardial infarction.

The frequencies of the HapA, HapB, and nine-marker haplotypes were not significantly different between the women of the control and MI groups and between the men of the two groups. In addition, we did not observe significant differences in age or sex between the carriers and noncarriers of specific haplotypes in the control group.

To assess whether independent associations existed between nine-marker haplotypes and MI, we performed a multivariate logistic regression analysis. After adjustments were made for conventional cardiovascular risk markers (age, gender, history of arterial hypertension, history of hypercholesterolemia, cur-

Table 6

Frequencies of nine-marker haplotypes in the control and MI groups

Haplotype		Control group	MI group	
Name	Allele combination	(2422 haplotypes)	(7314 haplotypes)	*P*
Hap1	GGAGTGCAG	928 (38.3)	2805 (38.4)	0.98
Hap2	GGAGTGAAG	237 (9.8)	808 (11.0)	0.082
Hap3	AGAGTGAAG	253 (10.4)	705 (9.6)	0.25
Hap4	GGGAAGAAG	204 (8.4)	665 (9.1)	0.32
Hap5	GAGAAGAAA	188 (7.8)	479 (6.5)	0.041
	Other	612 (25.3)	1852 (25.3)	0.96

Haplotype frequencies are presented as number (%). Shown are results obtained from the five most frequent nine-marker haplotypes and the combined other nine-marker haplotypes. Each haplotype is defined as a specific allele combination based on nine single nucleotide polymorphisms (SNPs) in *ALOX5AP*. The order of the SNPs is as follows (from left to right): SG13S25, SG13S377, SG13S100, SG13S106, SG13S114, SG13S89, SG13S32, SG13S41, SG13S35. Overall $P = 0.12$. See Table 1 and Figure 1 for the locations of the SNPs in the *ALOX5AP* genomic region.
MI, myocardial infarction.

rent cigarette smoking, diabetes mellitus), the analysis showed that none of the five most frequent nine-marker haplotypes, including the Hap5 haplotype, or the combined other haplotypes were significantly related with MI ($P \geq 0.38$).

DISCUSSION

The present data show that specific SNPs in *ALOX5AP* are not associated with MI in a large German population. Analyses in women and men did not reveal sex-specific relationships between the SNPs and MI. Specific four-marker haplotypes, the HapA and HapB haplotypes, and nine-marker haplotypes were not associated with MI. Most SNPs were in strong LD and the LD block structure was similar to those in other white populations.[4,14] Three of the nine SNPs examined here, the SG13S100, SG13S106, and SG13S114, were found to be significantly associated with MI in a population from Iceland that consisted of 779 unrelated patients with MI and 624 population-based control individuals.[1] However, significant associations between these SNPs and MI were not observed when adjustments were made for the number of markers tested.[1] None of the three SNPs was associated with MI in the present population. The HapA haplotype was associated with a twofold greater risk of MI in the Icelandic population (nominal $P = 0.0000023$; adjusted $P = 0.005$) but not in a sample from the UK (753 patients with MI and 730 control individuals).[1] In the same UK population, the HapB haplotype was associated with MI (nominal $P = 0.00037$; adjusted $P = 0.046$).[1]

The control subjects had some indication for coronary angiography, and, therefore, they did not constitute a typical sample of healthy controls. We compared the frequencies of the HapA haplotype and the SNP alleles that define the HapA haplotype between the present control sample and an independent control group that consisted of 736 unrelated individuals from the KORAS2000 sample, a representative local population sample from southern Germany.[4] In the present control group and the control group from the KORAS2000 sample,[4] the frequencies of the HapA haplotype were 14.8% versus 15.2% ($P = 0.74$) and the frequencies of the SG13S25-G, SG13S114-T, SG13S89-G, and SG13S32-A alleles were 89.3% vs. 90.1% ($P = 0.42$), 65.5% vs. 65.0% ($P = 0.77$), 95.0% vs. 96.0% ($P = 0.15$), and 50.5% vs. 49.7% ($P = 0.62$), respectively. Thus, with regard to the frequencies of the HapA haplotype and the alleles that constitute the HapA haplotype, the present control group is not substantially different from an established population-based sample. We inferred from this finding that the control sample with coronary angiography was suitable for the genetic association study described here. Measures of inflammation were not examined, which is a limitation of the current study.

Relationships of *ALOX5AP* SNPs and haplotypes with MI and ischemic stroke were evaluated in a nested case-control study within the Physicians' Health Study cohort that comprised predominantly white (>94%) male US physicians.[14,20] Investigation of 341 MI case-control pairs did not provide evidence of an association of any of the tested SNPs or the HapA or HapB haplotype with MI.[14] Genotype distributions and frequencies of SNP alleles and the HapA and HapB haplotypes in the case and control groups of the US sample[14] corresponded well with those of the present German sample.

Similar to results obtained in Germans (this study) and US physicians,[14] the SNPs that define the HapA and HapB haplotypes were not associated with MI in a Japanese population that included 353 patients with MI and 1875 control individuals.[2] A meaningful association analysis of the HapA and HapB haplotypes was not possible in the sample from Japan because, with some of the SNPs, minor alleles were either absent or extremely rare.[2] Two-marker *ALOX5AP* haplotypes not related to the HapA and HapB haplotypes were associated with MI in the Japanese sample.[2]

Studies conducted with samples of white individuals provided heterogeneous results about the relationship of the HapA and HapB haplotypes with MI (Table 7). Association of the HapA haplotype with MI was observed in a study sample from Iceland, but this finding was not replicated in samples from Germany (present study), the UK, and the US.[1,14] A relationship of the HapB haplotype with MI was found in a study sample from the UK, but this result was not confirmed in samples from Germany (present study) and the US.[1,14] Heterogeneities of genetic and environmental factors across the source populations are unlikely to account for the inconsistencies. Genetic markers for proposed gene-disease associations may vary in frequency between populations, but there is empirical evidence that their biological impact on the risk of common diseases is

Table 7

Frequencies of the HapA and Hap B haplotypes and estimated risks of MI associated with these haplotypes in case-control studies conducted with white population samples

	HapA			HapB		
Study	Controls/cases	Risk	P	Controls/cases	Risk	P
Germany (present)	0.15/0.16	1.10	0.16	0.08/0.07	0.94	0.48
United States[14]	0.14/0.17	1.18	0.46	0.07/0.06	0.62	0.08
Iceland[1]	0.10/0.16	1.80	<0.005[a]		No data available	
United Kingdom [1]	0.15/0.17	n.s.		0.04/0.08	1.95	0.046[a]

Haplotype frequencies are presented as proportions of controls and cases; n.s. not significant (data not shown).[1]

[a] Adjusted for the number of haplotypes tested.[1]

usually consistent even across ethnic boundaries.[21] Consistent replication of genetic associations has been difficult to achieve, despite the biological plausibility of these associations.[22] In this context, the present findings argue against association of defined SNPs and haplotypes of *ALOX5AP* with MI.

ACKNOWLEDGMENTS

The work was entirely supported by institutional financing from the German Heart Center Munich. The authors thank Wolfgang Latz, Marianne Eichinger, and Claudia Ganser for excellent technical assistance.

References

1. Helgadottir A, Manolescu A, Thorleifsson G, Gretarsdottir S, et al. The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nat Genet* 2004;36:233–239.
2. Kajimoto K, Shioji K, Ishida C, Iwanaga Y, et al. Validation of the association between the gene encoding 5-lipoxygenase-activating protein and myocardial infarction in a Japanese population. *Circ J* 2005;69:1029–1034.
3. Helgadottir A, Gretarsdottir S, St. Clair D, Manolescu A, et al. Association between the gene encoding 5-lipoxygenase-activating protein and stroke replicated in a Scottish population. *Am J Hum Genet* 2005;76:505–509.
4. Lõhmussaar E, Gschwendtner A, Mueller JC, Org T, et al. *ALOX5AP* gene and the *PDE4D* gene in a Central European population of stroke patients. *Stroke* 2005;36:731–736.
5. Miller DK, Gillard JW, Vickers PJ, Sadowski S, et al. Identification and isolation of a membrane protein necessary for leukotriene production. *Nature* 1990;343:278–281.
6. Dixon RAF, Diehl RE, Opas E, Rands E, et al. Requirement of a 5-lipoxygenase-activating protein for leukotriene synthesis. *Nature* 1990;343:282–284.
7. Libby P, Ridker PM, Maseri A. Inflammation and atherosclerosis. *Circulation* 2002;105:1135–1143.
8. Dwyer JH, Allayee H, Dwyer KM, Fan J, et al. Arachidonate 5-lipoxygenase promoter genotype, dietary arachidonic acid, and atherosclerosis. *N Engl J Med* 2004;350:29–37.
9. De Caterina R, Zampolli A. From asthma to atherosclerosis - 5-lipoxygenase, leukotrienes, and inflammation. *N Engl J Med* 2004;350:4–7.
10. Spanbroek R, Gräbner R, Lötzer K, Hildner M, et al. Expanding expression of the 5-lipoxygenase pathway within the arterial wall during human atherogenesis. *Proc Natl Acad Sci U S A* 2003;100:1238–1243.
11. Zhao L, Funk CD. Lipoxygenase pathways in atherogenesis. *Trends Cardiovasc Med* 2004;14:191–195.
12. Kennedy BP, Diehl RE, Boie Y, Adam M, et al. Gene characterization and promoter analysis of the human 5-lipoxygenase-activating protein (FLAP). *J Biol Chem* 1991;266:8511–8516.
13. Yandava CN, Kennedy BP, Pillari A, Duncan AM, et al. Cytogenetic and radiation hybrid mapping of human arachidonate 5-lipoxygenase-activating protein (ALOX5AP) to chromosome 13q12. *Genomics* 1999;56:131–133.
14. Zee RYL, Cheng S, Hegener HH, Erlich HA, et al. Genetic variants of arachidonate 5-lipoxygenase-activating protein, and risk of incident myocardial infarction and ischemic stroke. A nested case-control approach. *Stroke* 2006;37:2007–2011.
15. World Medical Association declaration of Helsinki. Recommendations guiding physicians in biomedical research involving human subjects. *JAMA* 1997;277:925–926.
16. Chalmers J, MacMahon S, Mancia G, Whitworth J, et al. 1999 World Health Organization-International Society of Hypertension Guidelines for the management of hypertension. Guidelines sub-committee of the World Health Organization. *Clin Exp Hypertens* 1999;21:1009–1060.
17. Diabetes mellitus. World Health Organization Study Group. Diabetes mellitus. *WHO Tech Rep Ser* 1985;727:1–104.
18. Kutyavin IV, Afonina IA, Mills A, Gorn VV, et al. 3'-Minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. *Nucleic Acids Res* 2000;28:655–661.
19. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001;68:978–989.
20. Final report on the aspirin component of the ongoing Physicians' Health Study. Steering committee of the Physicians' Health Study research group. *N Engl J Med* 1989;321:129–135.
21. Ioannidis JPA, Ntzani EE, Trikalinos TA. 'Racial' differences in genetic effects for complex diseases. *Nat Genet* 2004;36:1312–1318.
22. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002;4:45–61.

Nonvalidation of Reported Genetic Risk Factors for Acute Coronary Syndrome in a Large-Scale Replication Study

Thomas M. Morgan, MD
Harlan M. Krumholz, MD, MS
Richard P. Lifton, MD, PhD
John A. Spertus, MD, MPH

Context Given the numerous, yet inconsistent, reports of genetic variants being associated with acute coronary syndromes (ACS), there is a need for comprehensive validation of ACS susceptibility genotypes.

Objective To perform an extensive validation of putative genetic risk factors for ACS.

Design, Setting, and Participants Through a systematic literature search of articles published before March 10, 2005, we identified genetic variants previously reported as significant susceptibility factors for atherosclerosis or ACS. Restricting our analysis to white patients to reduce confounding from racial admixture, we identifed 811 patients who presented from March 2001 through June 2003 with ACS at 2 Kansas City, Mo, university-affiliated hospitals. During 2005-2006, we genotyped the 811 patients along with 650 age- and sex-matched controls for 85 variants in 70 genes and attempted to replicate previously reported associations. We further explored possible associations without prior assumption of specific risk models and used the Sign test to search for weak associations.

Main Outcome Measures Compare each prespecified gene variant associated with ACS risk among cases and controls. A surplus of associations would imply that some are associated with ACS.

Results Of 85 variants tested, only 1 putative risk genotype (−455 promoter variant in β-fibrinogen) was nominally statistically significant ($P=.03$). Only 4 additional genes were positive in model-free analysis. Neither number of associations was more frequent than expected by chance, given the number of comparisons. Finally, only 41 of 84 predefined risk variants were even marginally more frequent in cases than in controls (with 1 tie), representing a 48.8% "win rate" (95% confidence interval, 38.1%-59.5%) for the collective risk genotypes ($P=.91$, Sign test).

Conclusions Our null results provide no support for the hypothesis that any of the 85 genetic variants tested is a susceptibility factor for ACS. These results emphasize the need for robust replication of putative genetic risk factors before their introduction into clinical care.

JAMA. 2007;297:1551-1561 www.jama.com

COMPELLING EVIDENCE FROM twin and epidemiological studies suggests a genetic basis for atherosclerotic heart disease and acute coronary syndromes (ACS), including unstable angina, non–ST-elevation myocardial infarction (NSTEMI), and ST-elevation myocardial infarction (STEMI).[1,2] To date, numerous candidate genes have been implicated, mainly by case-control studies, as potential cardiovascular risk factors, but few, if any, have been established definitively.[3-5] Factors undermining the validity of previous reports include inappropriately small sample sizes, multiple subgroup comparisons, and publication bias.[4]

Before use in clinical care, potential genetic risk factors would ideally be replicated en masse in large, well-characterized patient populations.[6] To date, no such comprehensive validation of genetic variants potentially associated with ACS or atherosclerosis has been reported.

Accordingly, we first sought to identify genetic associations with ACS by systematically searching the medical literature for variants reported in association with MI, unstable angina, or atherosclerosis. We then attempted to validate these putative genetic risks in a large case-control study.

METHODS

Candidate Genes

We searched PubMed and bibliographies of original and review articles for manuscripts published before March 10, 2005, that reported statistically significant associations between specific genotypes and coronary atherosclerosis or ACS (A list of the articles is available on request from the authors). MEDLINE search terms

Author Affiliations: Department of Genetics, Howard Hughes Medical Institute (Drs Morgan and Lifton), Robert Wood Johnson Clinical Scholars Program and Department of Internal Medicine (Dr Krumholz), Yale University School of Medicine, New Haven, Conn, and Mid-America Heart Institute and University of Missouri-Kansas City, Mo (Dr Spertus). Dr Morgan is now with the Department of Pediatrics, Division of Genetics and Genomic Medicine, Washington University School of Medicine, St Louis, Mo.

Corresponding Author: Thomas M. Morgan, MD, Washington University School of Medicine, McDonnell Pediatric Research Bldg, 3103, 660 Euclid Ave, St Louis, MO 63110, (email: morgan_t@kids.wustl.edu) or Richard P. Lifton, MD, PhD, Yale University School of Medicine, 295 Congress Ave, New Haven, CT 06510 (richard.lifton@yale.edu).

included: *gene*, *genetic*, *polymorphism*, *myocardial infarction*, *atherosclerosis*, *coronary heart disease*, and *coronary artery disease*. Reports were included if they contained a claim of a significant positive association, with an investigator-reported *P* value <.05. A total of 96 polymorphic genetic variants in 75 genes were identified and included (TABLE 1 and TABLE 2). Eleven of those were excluded because they had failed the multiplex genotyping assay.

Description of Cases and Controls

Eight hundred eleven white patients of European ancestry with ACS were identified from a consecutive series of patients presenting at 2 Kansas City, Mo, hospitals (Mid-America Heart Institute and Truman Medical Center), from March 2001 through June 2003. Standard definitions were used to diagnose ACS patients with either MI or unstable angina.[92,93] Myocardial infarction was defined by a positive troponin blood test in the setting of symptoms and electrocardiogram changes (both ST-segment elevation and non–ST-segment elevation changes) consistent with MI. Unstable angina diagnoses were confirmed, by concurrence of 3 physician chart reviewers, if patients had negative troponin blood tests and any one of the following: new onset angina (<2 months) of at least Canadian Cardiovascular Society Classification class III, prolonged (>20 minutes) rest angina, recent (<2 months) worsening of angina, or angina that occurred within 2 weeks of an MI.[93] Of the troponin-negative unstable angina patients, 203 (92.7%) had a cardiac catheterization, a nuclear stress test, or a stress echocardiogram to corroborate their diagnoses.

Each participating inpatient with ACS was interviewed to determine variables, such as smoking, alcohol use, family history (≥1 first-degree relatives with MI or coronary artery disease), and to obtain consent for a blood sample for genetic analysis. In addition, detailed chart abstractions were performed to collect relevant laboratory and clinical data.

A total of 1045 ACS patients (of which 811 white patients were included in the current study) agreed to participate and to provide a blood sample for genetic analysis. Patients self-reported their race/ethnicity by selecting one of the following descriptors that were provided by the investigators: white, white Hispanic, African American, and African American non-Hispanic. Age- and sex-matched controls were recruited from the ambulatory outpatient clinical laboratory of 1 of the centers, Saint Luke's Hospital of Kansas City. These patients were undergoing routine laboratory testing and were asked to complete a medical questionnaire defining cardiac risk factors and medical comorbidities. Those controls reporting a previous ACS, prior coronary artery bypass graft surgery or prior percutaneous coronary intervention were excluded. To minimize the potential impact of genetic admixture, 650 white controls of mixed European ancestry who reported no history of coronary artery disease were selected from among the 1054 potential controls. Risk factor data were missing for 9 sex-, age-, and race-matched unaffected controls, and 56 additional matched controls were used for *ALOX5AP* haplotyping.

The research protocol was approved by the institutional review boards of both institutions; all study participants provided written informed consent for clinical and genetic studies.

Genotyping

Genomic DNA was isolated (Gentra PUREGENE, Minneapolis, Minn) from blood samples and subjected to whole genome amplification by multiple-strand displacement (Molecular Staging Inc, New Haven, Conn), using random priming and Phi-29 polymerase.[94,95] Genotyping was performed using the Sequenom MALDI-TOF (Matrix Assisted Laser Desorption-Ionization Time-of-Flight) system, using Spectrodesign software for assay design (Sequenom, San Diego, Calif), and assay methods that have previously been described.[96,97] Gene variants were excluded from analysis if they could not be genotyped using the Sequenom system due to persistent assay failure, defined as less than 95% scorable genotypes after 4 multiplex reaction design cycles. Eleven assays were ultimately excluded.* For the rare *MEF2A* 21–base pair (bp) deletion, cases and controls were genotyped by polymerase chain reaction to generate amplicons of 152-bp nondeletion or 131-bp deletion followed by electrophoresis on 3% agarose gels. Identified deletions were confirmed by direct DNA sequencing. Due to its rarity, *MEF2A* was analyzed separately, and thus only the other 84 genes were subjected to the full set of statistical analyses. PHASE Version 2.1 was used to estimate haplotype frequencies for *ALOX5AP*.[101,102]

Statistical Analysis

Genotype distributions in cases and controls were examined for significant deviation (*P*<.05) from Hardy-Weinberg equilibrium. The number of departures was assessed by Monte Carlo simulation and compared with the number expected by chance alone (Resampling Stats Inc, College Park, Md).

In the primary analysis, each genetic variant was prespecified based on published reports, and the frequencies of risk-associated variants were compared in cases and controls by using a 100 000 iteration Monte Carlo extension of the χ^2 test (SPSS 13.0 Exact Tests, SPSS Inc, Chicago, Ill). The term *statistically significant* was reserved for a *P* value below the Bonferroni-corrected study-wide significance threshold (0.05/84=0.0006). Because the Bonferroni correction is conservative when applied to a replication study, the total number of all positive associations at the *P*<.05 level was also compared with the expected number by chance in 100 000 simulations. A

*References 13, 20, 42, 45, 72, 88, 98-100.

Table 1. Validation of Predefined Risk Genotype Comparisons in Cases vs Controls

Gene Symbol	Variant	Genotype Comparison	Risk Variant Control Frequency	Odds Ratio (95% CI)	2-Tailed *P* Value	Genotype Frequency Difference
ABCA1	−477C/T[7,8]	TT vs CT or CC	0.222	1.13 (0.88-1.45)	.35	0.0215
ABCA1	Lys219Arg[9,10]	A vs G	0.274	1.02 (0.87-1.20)	.83	0.0041
ACE1	indel[4]	DD vs DI or II	0.286	1.07 (0.85-1.35)	.60	0.0139
ADD1	Gly460Trp[11,12]	T vs G	0.199	1.09 (0.91-1.32)	.35	0.0148
ADRB2§	Glu27Gln[13]	G vs C	0.441	0.93 (0.80-1.08)	.34	−0.0186
ADRB2	Ile164Thr[13]	CC vs CT	0.989	0.47 (0.20-1.13)	.10	−0.0117
ADRB2	Gly16Arg[13]	A vs G	0.377	1.04 (0.89-1.20)	.67	0.0082
ADRB3	Arg64Trp[14]	C vs T	0.077	0.99 (0.76-1.31)	>.99	−0.0004
AGT	Thr235Met[15]	T vs C	0.572	1.04 (0.89-1.20)	.65	0.0086
AGTR1	A1166C[16]	CC vs CA or AA	0.092	1.08 (0.76-1.53)	.72	0.0063
ALOX5AP	HAP B[17]	HAP B vs non-B	0.062	1.12 (0.90-1.40)	.31	0.0120
ALOX5AP	HAP A[17]	HAP A vs non-A	0.165	0.90 (0.75-1.10)	.32	−0.0130
APOA1	C83T[18]	T vs C	0.173	0.99 (0.81-1.20)	.92	−0.0019
APOA1	−75G/A[19]	A vs G	0.004	1.95 (0.68-5.54)	.23	0.0037
APOE	Arg158Cys[20]	CC vs CT or TT	0.023	1.61 (0.85-3.02)	.17	0.0134
APOE	−219T/G[21]	T vs G	0.475	1.14 (0.99-1.32)	.08	0.0329
BDKRB2	−58C/T[22]	C vs T	0.590	0.93 (0.80-1.08)	.37	−0.0169
CCL11†	Thr23Ala[23]	C vs T	0.837	0.97 (0.80-1.19)	.80	−0.0036
CCR2	Val64Ile[24]	GG vs AG or AA	0.855	0.94 (0.70-1.25)	.71	−0.0082
CCR5	Indel[25]	I vs D	0.907	0.79 (0.62-1.00)	.05	−0.0224
CD14	−159C/T[26]	TT vs CT or CC	0.237	1.10 (0.86-1.39)	.46	0.0170
CETP	intron1 G/A[27]	G vs A	0.568	0.93 (0.81-1.08)	.37	−0.0169
CETP	−629C/A[28]	C vs A	0.505	0.96 (0.83-1.11)	.60	−0.0104
COMT†	Val158Met[29]	GG or AG vs AA	0.222	1.11 (0.87-1.43)	.41	0.0193
CX3CR1	Ile249Val[30,31]	C vs T	0.729	0.96 (0.81-1.13)	.62	−0.0088
CX3CR1	Thr280Met[30]	G vs A	0.831	1.06 (0.87-1.29)	.58	0.0080
CYP11B2†	−344T/C[32,33]	C vs T	0.409	1.09 (0.94-1.26)	.27	0.0204
CYP2C9	Leu359Ile[34,35]	AC vs AA	0.094	0.76 (0.52-1.12)	.17	−0.0208
*CYP2C9**†	Cys144Arg[34,35]	CC vs CT	0.798	1.01 (0.77-1.32)	.95	0.0019
ENPP1	Gln121Lys[36]	C vs A	0.130	1.07 (0.86-1.32)	.59	0.0076
ESR1	−401T/C[37,38]	TT vs CT or CC	0.285	1.06 (0.84-1.33)	.64	0.0118
F12	46C/T[39]	TT vs CT or CC	0.067	0.92 (0.60-1.40)	.75	−0.0051
F13A1	Val34Leu[40]	G vs T	0.753	1.02 (0.86-1.22)	.79	0.0045
F2	G20210A[40,41]	A vs G	0.017	0.92 (0.52-1.64)	.88	−0.0013
F5	Arg506Gln[40]	A vs G	0.025	0.93 (0.58-1.50)	.81	−0.0016
F7†	Arg353Gln[42]	G vs A	0.893	0.94 (0.74-1.19)	.63	−0.0062
FGB	−455A/G[43]	GG or AG vs AA	0.607	1.27 (1.03-1.58)	.03	0.0558
GJA4	C1019T[44]	T vs C	0.318	0.90 (0.77-1.06)	.22	−0.0215
GP1BA	−5T/C[44-46]	T vs C	0.882	0.94 (0.75-1.17)	.57	−0.0071
GRL	Asn363Ser[47]	AG vs AA	0.073	0.79 (0.52-1.19)	.29	−0.0146
HFE	Cys282Tyr[48]	A vs G	0.065	0.98 (0.73-1.32)	.94	−0.0010
*HTR2A**	Ser102Ser[49]	TT vs CT or CC	0.156	1.12 (0.84-1.48)	.47	0.0154
ICAM1	Lys469Glu[50]	A vs G	0.557	1.09 (0.94-1.27)	.26	0.0214
IL1B	−511C/T[51]	CC vs CT or TT	0.445	1.14 (0.92-1.41)	.24	0.0328
IL6	−174G/C[52,53]	C vs G	0.403	1.05 (0.91-1.22)	.50	0.0127
IRS1	Arg971Gly[54]	A vs G	0.059	0.96 (0.70-1.32)	.81	−0.0023
ITGA2	Phe807Phe[55,56]	A vs G	0.391	1.03 (0.88-1.19)	.73	0.0063
ITGB3	Leu33Pro[4]	C vs T	0.164	0.85 (0.70-1.05)	.14	−0.0203
LIPC	−514T/C[57,58]	T vs C	0.240	0.85 (0.71-1.01)	.07	−0.0286
LPA	Asp9Asn[59]	AG vs GG	0.026	1.43 (0.78-2.63)	.29	0.0107
LRP1	Thr3261Thr[60]	GG or AG vs AA	0.107	1.02 (0.73-1.42)	.93	0.0018

(continued)

Table 1. Validation of Predefined Risk Genotype Comparisons in Cases vs Controls (cont)

Gene Symbol	Variant	Genotype Comparison	Risk Variant Control Frequency	Odds Ratio (95% CI)	2-Tailed *P* Value	Genotype Frequency Difference
LTA	A252G[61,62]	GG vs AG or AA	0.116	0.86 (0.62-1.20)	.40	−0.0147
LTA	Thr26Asn[61,62]	AA vs AC or CC	0.119	0.82 (0.59-1.14)	.27	−0.0191
MGP	Thr83Ala[63]	G vs A	0.385	1.00 (0.86-1.16)	>.99	0.0000
MGP	−7A/G[63]	G vs A	0.636	1.00 (0.86-1.16)	.97	−0.0010
MMP3	indel[64,65]	DD vs DI or II	0.284	0.82 (0.65-1.05)	.13	−0.0375
*MTHFR**	Ala222Val[66]	TT vs CT or CC	0.100	1.32 (0.95-1.84)	.10	0.0283
MTP	−493G/T[67,68]	T vs G	0.255	1.01 (0.86-1.20)	.86	0.0028
MTR	Asp919Gly[69]	A vs G	0.804	1.03 (0.85-1.24)	.78	0.0043
NPPA	Ter29ArgArg[70]	CC vs CT or TT	0.023	1.20 (0.61-2.32)	.62	0.0044
OLR1	Lys167Asn[71]	C vs G	0.916	0.82 (0.63-1.05)	.12	−0.0169
p22-PHOX‡	His72Tyr[72,73]	CC vs CT or TT	0.337	1.13 (0.91-1.39)	.28	0.0299
PAI1	Indel[43,44]	DD vs DI or II	0.309	1.00 (0.80-1.25)	>.99	−0.0005
PECAM1	Leu125Val[74,75]	GG vs CG or CC	0.288	0.94 (0.75-1.19)	.64	−0.0120
PECAM1	Ser563Asn[74,75]	A vs G	0.498	0.97 (0.84-1.13)	.71	−0.0072
PON1	Gln192Arg[76]	A vs G	0.705	1.01 (0.86-1.18)	.97	0.0010
PON2	Cys311Ser[77]	CC vs CG or GG	0.556	1.10 (0.89-1.35)	.40	0.0223
PPARG	Ala12Pro[78]	C vs G	0.129	0.83 (0.66-1.03)	.10	−0.0200
PTGS2	−765G/C[79]	C vs G	0.168	0.85 (0.69-1.04)	.11	−0.0219
RECQL2	Arg1367Cys[80]	T vs C	0.758	0.80 (0.68-0.94)	.01	−0.0433
SELE	Leu554Phe[75]	T vs C	0.036	1.17 (0.80-1.71)	.44	0.0059
SELE	Ser128Arg[75]	C vs A	0.103	0.91 (0.71-1.16)	.45	−0.0086
SELP	Thr715Pro[81]	A vs C	0.902	0.93 (0.73-1.18)	.58	−0.0068
TFPI	Val264Met[82]	AG vs GG	0.050	0.80 (0.49-1.32)	.44	−0.0096
THBD	−33G/A[83]	AG vs GG	0.002	2.39 (0.25-23.0)	.63	0.0022
THBD	Ala25Thr[84]	AG vs GG	0.011	1.29 (0.50-3.35)	.64	0.0030
THBD	Ala455Val[85]	CC vs CT or TT	0.659	1.05 (0.85-1.31)	.65	0.0114
THBS1	Asn700Ser[86]	GG vs AG or AA	0.021	0.81 (0.38-1.72)	.70	−0.0039
THBS2‡	3′UTR T/G[87]	TT or GT vs GG	0.950	0.51 (0.34-0.79)	.002	−0.0432
THBS4	Ala387Pro[87]	GG or CG vs CC	0.939	1.00 (0.65-1.53)	>.99	−0.0002
THPO	A5713G[88]	GG vs AG or AA	0.264	1.20 (0.95-1.51)	.13	0.0368
TLR4	Gly299Asp[89]	A vs G	0.942	1.02 (0.75-1.40)	.94	0.0011
TNF	−308G/A[90]	A vs G	0.158	0.86 (0.70-1.06)	.17	−0.0188
TNFRSF1A	Arg92Gln[91]	AG vs GG	0.050	0.80 (0.49-1.32)	.44	−0.0096

*Hardy-Weinberg equilibrium deviation in controls, $P<.05$ (n = 3).
†Hardy-Weinberg equilibrium deviation in cases, $P<.05$ (n = 5).
‡$P<.001$ (n = 2).
§$P<.001$ (n = 1).

surplus of positive associations over random expectations would imply that some are truly associated with ACS.

Secondarily, we also compared the overall genotype distributions at each locus in cases and controls by Monte Carlo χ^2 testing. Power to confirm individual genetic associations was determined using a log-likelihood-based method (Quanto 1.0).[103,104]

Finally, as a measure to increase power, the observed proportion of prespecified risk variants found to be even marginally more frequent in cases than in controls was assessed by the Sign test. Under the null hypothesis, each of the risk variants is equally likely to be more frequent in cases, or in controls. To estimate the Sign test's power to detect an excess of even weakly positive genetic associations (50 of 84 positive associations confers $P=.05$ in the Sign test), we simulated the resampling of 650 control and 811 case genotypes across 84 genetic comparisons, finding the minimum detectable odds ratio ensuring a critical probability level of a 63.3% win rate for each 84 risk variants that provides 80% confidence of having at least 50 wins.

RESULTS

The clinical characteristics of the 811 cases and 650 controls are described in TABLE 3 and the distributions of their genotypes are shown in Table 2. The population of ACS cases included 308 (38%) STEMI, 284 (35%) NSTEMI, and 219 (27%) unstable angina patients. Cases and controls had similar age, sex, and body mass index distributions. A family history of coronary artery disease or MI among first-degree relatives was 2.7-fold higher in male cases than in male controls and 2.0-fold higher in female cases than in female controls.

Table 2. Genotype Frequencies and *P* Values in Cases With Acute Coronary Syndrome and Controls

			No. (%)		
Gene	Variant	Genotype	Cases	Controls	2-Tailed *P* Value
ABCA1	−477C/T	CC	191 (24.6)	182 (28.3)	.27
		CT	396 (51.0)	319 (49.5)	
		TT	189 (24.4)	143 (22.2)	
ABCA1	Lys219Arg	AA	65 (8.2)	46 (7.1)	.69
		AG	311 (39.3)	263 (40.6)	
		GG	416 (52.5)	338 (52.2)	
ACE1	I/D	DD	233 (30.0)	185 (28.6)	.84
		DI	389 (50.1)	329 (50.9)	
		II	154 (19.8)	132 (20.4)	
ADD1	Gly460Trp	GG	456 (60.7)	419 (64.9)	.08
		GT	269 (35.8)	197 (30.5)	
		TT	26 (3.5)	30 (4.6)	
ADRB2	Glu27Gln†	CC	266 (34.5)	217 (34.8)	.14
		CG	358 (46.5)	264 (42.3)	
		GG	146 (19.0)	143 (22.9)	
ADRB2	Ile164Thr	CC	789 (97.8)	652 (98.9)	.11
		CT	18 (2.2)	7 (1.1)	
		TT	0	0	
ADRB2	Gly16Arg	AA	128 (16.3)	100 (15.2)	.87
		AG	348 (44.3)	294 (44.8)	
		GG	309 (39.4)	262 (39.9)	
ADRB3	Arg64Trp	CC	6 (0.7)	1 (0.2)	.21
		CT	111 (13.8)	99 (15.1)	
		TT	687 (85.4)	557 (84.8)	
AGT	Thr235Met	CC	143 (17.8)	107 (16.5)	.27
		CT	387 (48.3)	340 (52.6)	
		TT	272 (33.9)	200 (30.9)	
AGTR1	A1166C	AA	388 (48.1)	332 (50.8)	.61
		AC	339 (42.1)	262 (40.1)	
		CC	79 (9.8)	60 (9.2)	
ALOX5AP	HAP B	B	50 (6.5)	41 (5.8)	.31
		non-B	734 (93.5)	661 (94.2)	
			NA	NA	
ALOX5AP	HAP A	A	124 (15.9)	122 (17.4)	.35
		non-A	654 (84.1)	584 (82.6)	
			NA	NA	
APOA1	C83T	AA	23 (3.0)	25 (3.8)	.59
		AG	219 (28.3)	175 (26.9)	
		GG	532 (68.7)	450 (69.2)	
APOA1	−75G/A	AA	1 (0.1)	0	.51
		AG	10 (1.3)	5 (0.8)	
		GG	784 (98.6)	638 (99.2)	
APOE	Arg158Cys	CC	29 (3.6)	15 (2.3)	.14
		CT	209 (26.1)	154 (23.5)	
		TT	562 (70.3)	487 (74.2)	
APOE	−219T/G	GG	194 (24.2)	177 (27.3)	.21
		GT	403 (50.2)	327 (50.5)	
		TT	206 (25.7)	144 (22.2)	
BDKRB2	−58C/T	CC	263 (32.8)	221 (33.6)	.43
		CT	394 (49.1)	335 (50.9)	
		TT	145 (18.1)	102 (15.5)	
CCL11	Thr23Ala‡	CC	539 (68.2)	456 (70.0)	.20
		CT	239 (30.3)	178 (27.3)	
		TT	12 (1.5)	17 (2.6)	
CCR2	Val64Ile	AA	7 (0.9)	6 (0.9)	.91
		AG	116 (14.4)	89 (13.6)	
		GG	681 (84.7)	561 (85.5)	
CCR5	Indel	II	631 (78.4)	540 (82.4)	.15
		ID	162 (20.1)	108 (16.5)	
		DD	12 (1.5)	7 (1.1)	
CD14	−159C/T	CC	204 (25.4)	193 (29.5)	.22
		CT	395 (49.2)	306 (46.8)	
		TT	204 (25.4)	155 (23.7)	
CETP	intron1 G/A	AA	168 (20.9)	135 (20.6)	.44
		AG	387 (48.1)	297 (45.3)	
		GG	250 (31.1)	224 (34.1)	
CETP	−629C/A	AA	205 (25.6)	171 (26.4)	.31
		AC	400 (49.9)	298 (46.1)	
		CC	197 (24.6)	178 (27.5)	
COMT	Val158Met‡	AA	231 (29.8)	181 (28.1)	.39
		AG	358 (46.1)	321 (49.8)	
		GG	187 (24.1)	143 (22.2)	
CX3CR1	Ile249Val	CC	410 (51.1)	353 (53.6)	.43
		CT	336 (41.9)	254 (38.6)	
		TT	56 (7.0)	51 (7.8)	
CX3CR1	Thr280Met	AA	18 (2.2)	13 (2.0)	.65
		AG	223 (27.7)	195 (29.8)	
		GG	565 (70.1)	447 (68.2)	
CYP11B2	−344T/C‡	CC	163 (20.6)	109 (16.6)	.12
		CT	352 (44.6)	319 (48.6)	
		TT	275 (34.8)	229 (34.9)	
CYP2C9	Leu359Ile	AA	708 (92.7)	568 (90.6)	.17
		AC	56 (7.3)	59 (9.4)	
		CC	0	0	
CYP2C9	Cys144Arg*‡	CC	589 (80.0)	491 (79.8)	.95
		CT	147 (20.0)	124 (20.2)	
		TT	0	0	
ENPP1	Gln121Lys	AA	600 (74.3)	498 (75.7)	.84
		AC	192 (23.8)	149 (22.6)	
		CC	15 (1.9)	11 (1.7)	
ESR1	−401T/C	CC	145 (18.0)	143 (21.8)	.20
		CT	421 (52.3)	326 (49.7)	
		TT	239 (29.7)	187 (28.5)	
F12	46C/T	CC	459 (58.0)	371 (56.6)	.84
		CT	283 (35.8)	241 (36.7)	
		TT	49 (6.2)	44 (6.7)	
F13A1	Val34Leu	GG	443 (56.8)	354 (55.8)	.93
		GT	296 (37.9)	247 (39.0)	
		TT	41 (5.3)	33 (5.2)	
F2	G20210A	AA	1 (0.1)	1 (0.2)	.94
		AG	23 (2.9)	20 (3.0)	
		GG	783 (97.0)	635 (96.8)	

(continued)

(continued)

Table 2. Genotype Frequencies and *P* Values in Cases With Acute Coronary Syndrome and Controls (cont)

Gene	Variant	Genotype	No. (%)		2-Tailed *P* Value
			Cases	Controls	
F5	Arg506Gln	AA	1 (0.1)	1 (0.2)	.95
		AG	36 (4.5)	31 (4.7)	
		GG	769 (95.4)	623 (95.1)	
F7	Arg353Gln‡	AA	16 (2.0)	6 (0.9)	.22
		AG	148 (18.7)	129 (19.6)	
		GG	629 (79.3)	522 (79.5)	
FGB	−455A/G	AA	24 (3.0)	26 (4.0)	.08
		AG	247 (30.7)	229 (35.3)	
		GG	533 (66.3)	394 (60.7)	
GJA4	C1019T	CC	401 (50.6)	311 (47.5)	.47
		CT	313 (39.5)	272 (41.5)	
		TT	78 (9.8)	72 (11.0)	
GP1BA	−5T/C	CC	13 (1.7)	8 (1.2)	.75
		CT	168 (21.6)	138 (21.1)	
		TT	597 (76.7)	509 (77.7)	
GRL	Asn363Ser	AA	756 (94.1)	608 (92.7)	.29
		AG	47 (5.9)	48 (7.3)	
		GG	0	0	
HFE	Cys282Tyr	AA	3 (0.4)	1 (0.2)	.70
		AG	96 (12.0)	83 (12.6)	
		GG	703 (87.7)	574 (87.2)	
HTR2A	Ser102Ser*	CC	286 (36.5)	275 (42.0)	.11
		CT	363 (46.4)	278 (42.4)	
		TT	134 (17.1)	102 (15.6)	
ICAM1	Lys469Glu	AA	270 (34.0)	195 (30.2)	.30
		AG	379 (47.7)	329 (51.0)	
		GG	145 (18.3)	121 (18.8)	
IL1B	−511C/T	CC	359 (47.7)	289 (44.5)	.40
		CT	311 (41.4)	292 (44.9)	
		TT	82 (10.9)	69 (10.6)	
IL6	−174G/C	CC	142 (17.6)	106 (16.1)	.73
		CG	386 (48.0)	319 (48.5)	
		GG	277 (34.4)	233 (35.4)	
IRS1	Arg971Gly	AA	3 (0.4)	3 (0.5)	.98
		AG	84 (10.6)	69 (10.9)	
		GG	704 (89.0)	562 (88.6)	
ITGA2	Phe807Phe	AA	123 (15.3)	108 (16.4)	.40
		AG	394 (48.9)	298 (45.4)	
		GG	288 (35.8)	251 (38.2)	
ITGB3	Leu33Pro	CC	20 (2.5)	14 (2.2)	.12
		CT	188 (23.6)	182 (28.3)	
		TT	588 (73.9)	446 (69.5)	
LIPC	−514T/C	CC	506 (62.9)	369 (56.8)	.03
		CT	256 (31.8)	250 (38.5)	
		TT	42 (5.2)	31 (4.8)	
LPA	Asp9Asn	AG	29 (3.7)	17 (2.6)	.29
		GG	765 (96.3)	641 (97.4)	
		GG	0	0	
LRP1	Thr3261Thr	AA	367 (46.9)	283 (43.2)	.31
		AG	330 (42.2)	302 (46.1)	
		GG	85 (10.9)	70 (10.7)	
LTA	A252G	AA	394 (49.1)	282 (429)	.06
		AG	327 (40.8)	299 (45.5)	
		GG	81 (10.1)	76 (11.6)	
LTA	Thr26Asn	AA	80 (10.0)	78 (11.6)	.07
		AC	331 (41.4)	297 (45.3)	
		CC	389 (48.6)	280 (42.7)	
MGP	Thr83Ala	AA	308 (38.3)	257 (39.1)	.79
		AG	374 (46.5)	294 (44.7)	
		GG	123 (15.3)	106 (16.1)	
MGP	−7A/G	AA	110 (13.6)	95 (14.5)	.77
		AG	368 (45.7)	288 (43.8)	
		GG	328 (40.7)	274 (41.7)	
MMP3	indel	DD	194 (24.7)	176 (28.4)	.27
		DI	386 (49.1)	294 (47.5)	
		II	206 (26.2)	149 (24.1)	
MTHFR	Ala222Val*	CC	350 (44.1)	272 (41.3)	.06
		CT	341 (43.0)	320 (48.6)	
		TT	102 (12.9)	66 (10.0)	
MTP	−493G/T	GG	449 (55.8)	371 (56.5)	.95
		GT	297 (36.9)	237 (36.1)	
		TT	59 (7.3)	49 (7.5)	
MTR	Asp919Gly	AA	529 (66.0)	423 (65.7)	.88
		AG	239 (29.8)	190 (29.5)	
		GG	34 (4.2)	31 (4.8)	
NPPA	Ter29ArgArg	CC	22 (2.8)	15 (2.3)	.83
		CT	190 (23.9)	159 (24.7)	
		TT	583 (73.3)	471 (73.0)	
OLR1	Lys167Asn	CC	649 (80.8)	543 (83.7)	.26
		CG	146 (18.2)	103 (15.9)	
		GG	8 (1.0)	3 (0.5)	
P22-PHOX	His72Tyr§	CC	347 (47.0)	288 (44.0)	.002
		CT	271 (36.7)	293 (44.7)	
		TT	121 (16.4)	74 (11.3)	
PAI1	indel	DD	249 (30.9)	203 (30.9)	.77
		DI	398 (49.4)	314 (47.9)	
		II	159 (19.7)	139 (21.2)	
PECAM1	Leu125Val	CC	187 (23.3)	155 (23.6)	.82
		CG	395 (49.1)	312 (47.6)	
		GG	222 (27.6)	189 (28.8)	
PECAM1	Ser563Asn	AA	200 (25.0)	163 (25.5)	.92
		AG	386 (48.3)	312 (48.8)	
		GG	214 (26.8)	165 (25.8)	
PON1	Gln192Arg	AA	396 (49.6)	324 (49.3)	>.99
		AG	337 (42.2)	279 (42.5)	
		GG	66 (8.3)	54 (8.2)	
PON2	Cys311Ser	CC	464 (57.9)	366 (55.6)	.50
		CG	298 (37.2)	251 (38.1)	
		GG	40 (5.0)	41 (6.2)	
PPARG	Ala12Pro	CC	637 (79.2)	492 (75.9)	.22
		CG	159 (19.8)	145 (22.4)	
		GG	8 (1.0)	11 (1.7)	

(continued)

(continued)

Table 2. Genotype Frequencies and *P* Values in Cases With Acute Coronary Syndrome and Controls (cont)

			No. (%)		
Gene	Variant	Genotype	Cases	Controls	2-Tailed *P* Value
PTGS2	−765G/C	CC	15 (1.9)	21 (3.3)	.17
		CG	202 (25.5)	174 (27.1)	
		GG	576 (72.6)	447 (69.6)	
RECQL2	Arg1367Cys	CC	66 (8.2)	38 (5.8)	.03
		CT	326 (40.5)	239 (36.7)	
		TT	412 (51.2)	375 (57.5)	
SELE	Leu554Phe	CC	740 (91.9)	611 (92.9)	.47
		CT	63 (7.8)	47 (7.1)	
		TT	2 (0.2)	0	
SELE	Ser128Arg	AA	658 (82.0)	528 (80.4)	.71
		AC	137 (17.1)	123 (18.7)	
		CC	7 (0.9)	6 (0.9)	
SELP	Thr715Pro	AA	646 (80.2)	530 (80.8)	.22
		AC	150 (18.6)	124 (18.9)	
		CC	9 (1.1)	2 (0.3)	
TFPI	Val264Met	AA	758 (95.9)	625 (95.0)	.45
		AG	32 (4.1)	33 (5.0)	
		GG	0	0	
THBD	−33G/A	AA	801 (99.6)	638 (99.8)	.63
		AG	3 (0.4)	1 (0.2)	
		GG	0	0	
THBD	Ala25Thr	AA	794 (98.6)	652 (98.9)	.64
		AG	11 (1.4)	7 (1.1)	
		GG	0	0	
THBD	Ala455Val	CC	531 (67.0)	433 (65.9)	.91
		CT	237 (29.9)	203 (3.9)	
		TT	24 (3.0)	21 (3.2)	
THBS1	Asn700Ser	AA	614 (76.3)	507 (77.2)	.74
		AG	177 (22.0)	136 (20.7)	
		GG	14 (1.7)	14 (2.1)	
THBS2	3'UTR T/G‡	GG	74 (9.4)	33 (5.0)	.001
		GT	250 (31.6)	251 (38.4)	
		TT	466 (59.0)	370 (56.6)	
THBS4	Ala387Pro	CC	49 (6.1)	40 (6.1)	.85
		CG	268 (33.4)	229 (34.8)	
		GG	486 (60.5)	389 (59.1)	
THPO	A5713G	AA	187 (23.3)	159 (24.2)	.30
		AG	374 (46.6)	324 (49.4)	
		GG	241 (30.0)	173 (26.4)	
TLR4	Gly299Asp	AA	702 (88.7)	579 (88.4)	.89
		AG	88 (11.1)	76 (11.6)	
		GG	1 (0.1)	0	
TNF	−308G/A	AA	17 (2.1)	14 (2.2)	.27
		AG	189 (23.5)	176 (27.2)	
		GG	597 (74.3)	457 (70.6)	
TNFRSF1A	Arg92Gln	AA	784 (97.0)	627 (95.4)	.13
		AG	24 (3.0)	30 (4.6)	
		GG	0	0	

*Hardy-Weinberg equilibrium deviation in controls, $P<.05$ (n = 3).
†$P<.001$ (n = 1).
‡Hardy-Weinberg equilibrium deviation in cases, $P<.05$ (n = 5).
§$P<.001$ (n = 2).

(continued)

Male and female cases were significantly more likely to be current smokers and to have type 2 diabetes mellitus but less likely to consume at least 1 alcoholic drink per month. Frequencies of hypercholesterolemia and hypertension were higher in female cases than in controls; no significant differences were observed in males. Previous revascularization had been performed in 35.6% of incident ACS cases and in none of the controls.

A total of 85 variants in 70 genes were genotyped in cases and controls.The overall genotype call rate for these variants was 98.5% (range, 95.0%-99.8%). Two percent of all samples were genotyped in duplicate for each marker in a blinded fashion as a measure of genotype reproducibility. Among the 2511 repeated genotypes, 5 were discordant, demonstrating a reproducibility of 99.8%.

Tests of Hardy-Weinberg equilibrium revealed that 1 variant violated it in both cases and controls, at the $P<.05$ level; 7 violated it in cases only; and 4 violated it in controls only (Table 1 and Table 2). This finding is not more than expected by chance (4 violations expected by chance in each group; see the Methods section) and therefore none was excluded from further analysis at this stage.

With respect to power parameters, the mean effective frequency (or 1-frequency, if $q>0.5$) in controls of the putative risk variants studied was 0.20, and 58 (68.2%) were common, (≥0.1), 25 (29.4%) were uncommon (<0.1; >0.01), and 2 (2.4%) were rare (≤0.01). Our sample had 80% power to confirm, by the Monte Carlo χ^2 test, a genotype-specific relative risk of 2.3 for a rare variant ($q=0.01$), 1.4 for a relatively uncommon variant ($q=0.1$), and 1.25 for a common allele ($q=0.5$).

We tested whether each putative risk variant showed a significant difference in frequency between cases and controls (Table 1). An odds ratio greater than 1 indicates that the risk genotype was in higher frequency among cases, and if so, the genotype frequency difference was reported as a positive decimal number. Only 1 genetic variant was significant at the $P<.05$ level, which is the number most likely by chance alone. The −455 variant, which lies upstream of the transcription initiation site in the β-fibrinogen gene, replicated the originally reported association, with the *GG* genotype being more frequent in cases than controls (frequency, 66% in cases vs 61% in controls; odds ratio, 1.27; $P=.03$). In addition, we found the *MEF2A* 21-bp deletion in 1 case and 1 control, con-

Table 3. Characteristics of 1461 White Participants Genotyped for 85 Genetic Variants*

	Men (n = 944)		Women (n = 517)	
Characteristics	ACS Cases (n = 550)	Controls (n = 394)	ACS Cases (n = 261)	Controls (n = 256)
Age, mean (SD), y	6.7 (12.5)	6.0 (12.1)	63.1 (13.2)	61.8 (12.8)
Body mass index, mean (SD)†	29.1 (5.5)	27.9 (5.0)	29.9 (6.9)	27.7 (6.9)
Family history of CAD/MI	279 (50.7)‡	109 (27.7)	135 (51.7)‡	90 (35.5)
Prior myocardial infarction	142 (25.8)‡	0	74 (28.4)‡	0
Prior revascularization	205 (37.3)‡	0	83 (31.8)‡	0
Congestive heart failure	23 (4.2)‡	0	18 (6.9)‡	0
Hypertension	305 (55.5)	207 (52.5)	182 (69.7)‡	126 (49.2)
Type 2 diabetes mellitus	116 (21.1)‡	42 (10.7)	77 (29.5)‡	35 (13.7)
Hypercholesterolemia	314 (57.1)	208 (52.8)	162 (62.1)‡	117 (45.7)
Postmenopausal			189 (68.6)‡	219 (85.5)
College graduate	166 (30.2)‡	238 (60.4)	40 (15.3)‡	72 (28.1)
Smoking <30 d ago	183 (33.3)‡	55 (14.0)	85 (32.6)‡	31 (12.1)
Alcohol frequency >1/mo	221 (40.2)‡	210 (53.3)	38 (14.6)‡	84 (32.8)

Abbreviations: ACS, acute coronary syndromes; CAD, coronary artery disease; MI, myocardial infarction.
*Data are presented as number (percentage) unless otherwise indicated.
†Body mass index is calculated as weight in kilograms divided by height in meters squared.
‡$P<.001$ for the comparison with controls of the same sex.

firming that this is a rare variant in the population.[105]

Several supplementary analyses were performed. When the genotypes of cases and controls were analyzed by extension of 2×3 χ^2 tests to 100 000 simulations, 4 loci, *RECQL2*, *THBS2*, *LIPC*, and *p22-PHOX*, were marginally significant (Table 2). In each case, the specific genetic risk model providing significance was different from that reported in the literature; hence, these cannot be considered formal replications and the total number of positive associations is not in excess of random expectations.

Finally, we found that only 41 of 84 predefined risk variants were even marginally more frequent in cases than in controls (excluding 1 tie, the rare *MEF2A* deletion), representing a 48.8% win rate (95% confidence interval, 38.1%-59.5%) for the collective-risk genotypes. This observed proportion of wins is not different from the expected proportion (50%) under the null hypothesis ($P=.91$). Table 1 shows that the absolute differences in risk genotype frequencies between cases and controls (negative signs meaning that the putative risk genotype was more frequent in controls than in cases) were small, with a median difference of −0.0003, and maximum of 0.056 (β fibrinogen).

COMMENT

We were unable to confirm as risk factors for ACS 85 genetic variants because none was unequivocally validated in this large case-control study of 1461 participants. In the primary analysis, only the −455 promoter variant in β-fibrinogen) was nominally statistically significant ($P=.03$). Among the 4 variants in the secondary analysis that met nominal statistical thresholds, there was an excess of a different variant than was previously reported among cases in the original study, which does not support replication. We therefore conclude that our findings, in this large sample of well-characterized ACS patients and controls, cannot support that this panel of gene variants contains bona fide ACS risk factors.

Our findings come at a critical juncture in complex disease genetics. Some cardiovascular gene variants (eg, *ACE*, *AGT*, *AGTR1*, *ITGB3*, *F2*, *F5*, *MTHFR*) included in our study can already be ordered clinically, for indications that explicitly include possible ACS risk. However, our findings suggest that such clinical genetic testing is premature and underscore the importance of robust replication studies of reported associations prior to their application to clinical care.

These nonreplications include variants in several high-profile studies. For example, haplotypes A and B of 5-lipoxygenase activating protein (*ALOX5AP*) were reported in 1 study to be associated with MI in the general populations of Iceland, and the United Kingdom, respectively.[17] We found neither haplotype was associated with ACS, in spite of our observed haplotype frequencies in cases and controls closely approximating those found in the total United Kingdom data set (cases and controls) previously (haplotype A, 0.165 vs 0.160, respectively; haplotype B, 0.062 vs 0.058).

Although our study raises significant doubts about the collective panel of putative genetic risk factors, it does not invalidate any particular previous study. Possible explanations of our negative results could include: (1) false-negative results in our study; (2) false-positive associations in previous studies; and (3) varied effects of risk variants in different genetic backgrounds.

False-negative results as a general explanation for our study's null findings are unlikely given that our sample size is substantially larger than all but a few reported prior studies and was powered to detect modest relative risks. Based on a random sample (n=30) of articles included in this study (1 per gene variant), we estimated that the mean odds ratio reported in positive studies was 2.3 (range, 1.25-5.0), indicating that we had well in excess of 80% power to replicate most reports. However, isolated positive reports may overestimate genetic risks.[5,6] Recently, a meta-analysis of 14 genes included in our study reported odds ratios ranging from 1.10 to 1.73 for risk of MI.[3] It is possible that minute odds ratios are to be expected in complex disease genetics and that neither our study nor most previous studies were sufficiently powered. Accordingly, we augmented our power, by use of the Sign test, to detect a surplus of as few as 16 weakly positive genetic risk factors among the entire set that we genotyped (84 −16=50, the number required for a significant Sign test), corresponding to a mean odds ratio of 1.05 or higher given our sample size and the average risk genotype frequency.

Absence of genetic effect only in our cohort is also unlikely. Cases showed a 2-fold higher family history of ACS, consistent with a genetic effect contributing to phenotypes in this cohort. In addition, homozygosity coding for an arginine residue at position 158 of apolipoprotein E (E4 variant), considered 1 of the least controversial of the putative ACS susceptibility factors despite some inconsistency in certain cohorts,[106] was significantly associated (P=.04) among cases with hyperlipidemia (4.1%) vs controls without hyperlipidemia (1.6%).

False-positive results in previous studies are another potential explanation for the discrepancy between our findings and those of others. This issue has previously been recognized as a serious problem with association studies, particularly when sample sizes are underpowered.[107] It is difficult to identify true vs false positives by analysis of the literature alone.[108] Unrecognized stratification between cases and controls can create spurious associations,[109] and the absence of negative genomic controls in nearly all prior studies to exclude this possibility leaves this an open question. Also difficult to assess is the extent to which publication bias and multiple hypothesis testing have had an effect.

It could be argued that our research participants are distinct from those reported previously and that our results may not bear on the validity of positive associations reported in different populations and clinical subgroups (eg, analyses substratified by age, sex, or a clinical variable, such as hypertension, hyperlipidemia, or smoking status). Given that the vast majority of common variants in the human genome date to our shared ancestry in Africa,[110] it is not likely that there are different common functional variants in linkage disequilibrium with risk variants in our population vs others. Less common mutations of more recent ancestral origin could conceivably be correlated with certain genetic variants in one population but not another. The extent to which linkage disequilibrium patterns might explain our findings is unknown, but our study population is quite typical of the mixed European background that is prevalent in the United States.

Another possibility is that the effect of risk variants is different in different genetic backgrounds; if true, the lack of generalizability of results will severely limit their application to the clinical arena. The fact that we failed to replicate positive associations in a consecutive series of study participants that are broadly representative of the disease encountered in clinical practice places limitations on the potential applicability of prior findings and supports our premise that it is premature to extrapolate these earlier findings to routine clinical care.

The failure of the candidate gene approach to identify variants conferring susceptibility to ACS risk prompts consideration of other approaches. One promising approach is to screen the entire genome in an unbiased way in a large sample for variants that are significantly associated with disease risk. Coupled with the understanding of underlying patterns of linkage disequilibrium in the human genome[7] and the ability to inexpensively obtain genotypes across the genome, the field is moving rapidly toward a comprehensive genome-wide approach. Challenges of this approach include the unknown number of variants that impart effect, the magnitude of the effect imparted by each, and the extent to which common variants as opposed to rare independent mutations account for disease risk.

Regardless of the approach taken, it is clear that multiple large, well-matched cohorts of cases and controls will be required to achieve valid progress in the genetic analysis of ACS and other complex human diseases. Our null findings indicate the need for caution in the interpretation of genetic associations in different clinical populations and the need for extensive validation of genetic risk factors.

Author Contributions: Dr Morgan had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Morgan, Lifton, Krumholz, Spertus.
Acquisition of data: Morgan, Lifton, Krumholz, Spertus.
Analysis and interpretation of data: Morgan, Lifton, Krumholz, Spertus.
Drafting of the manuscript: Morgan, Lifton, Krumholz, Spertus.
Critical revision of the manuscript for important intellectual content: Morgan, Lifton, Krumholz.
Statistical analysis: Morgan, Lifton.
Obtained funding: Morgan, Lifton, Spertus.
Administrative, technical, or material support: Lifton, Spertus.
Study supervision: Lifton, Krumholz, Spertus.

Financial Disclosures: Dr Spertus reports that he serves on the advisory boards of the American College of Cardiology, American Heart Association, Amgen United Healthcare, Blue Cross/Blue Shield; has received gants from the National Institutes of Health (NIH), Amgen, CV Therapeutics, Flowcardia, and Roache Diagnostics (in-kind biomarker reagent supplies for an NIH grant); has ownership interests in the Seattle Angina Questionnaire, the Kansas City Cardiomyopathy Questionnaire, the Peripheral Artery Questionnaire, and Health Outcomes Sciences; and has consulted within the past 5 years for CV Therapeutics, Amgen, Worldheart, and Ostuka Parmaceuticals. Dr Krumholz reports that he has research contracts with the Colorado Foundation for Medical Care and the American College of Cardiology, serves on the advisory boards for Amgen, Alere, and United Healthcare, is a subject-matter expert for VHA Inc. Drs Morgan and Lifton report no conflicts of interest.

Funding/Support: This project was funded by grants from the Saint Luke's Hospital Foundation, Kansas City, Mo, and by grant R-01 HS11282-01 from the Agency for Healthcare Research and Quality. Dr Morgan's research in Dr Lifton's laboratory at Yale University was supported by Howard Hughes Medical Institute and by grant NHLBI K23 HI77272, a mentored patient-oriented research grant from the National Heart, Lung, and Blood Institute.

Role of the Sponsor: None of the funding organizations had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

Acknowledgment: We thank Donna Buchanan, PhD, Mid-America Heart Institute, Kansas City, Mo,for editorial assistance with the manuscript as part of her duties and received no additional compensation.

REFERENCES

1. Marenberg ME, Risch N, Berkman LF, Floderus B, de Faire U. Genetic susceptibility to death from coronary heart disease in a study of twins. *N Engl J Med*. 1994;330:1041-1046.
2. Scheuner MT. Clinical application of genetic risk assessment strategies for coronary artery disease: genotypes, phenotypes, and family history. *Prim Care*. 2004;31:711-737, xi-xii.
3. Casas JPCJ, Miller GJ, Hingorani AD, Humphries SE. Investigating the genetic determinants of cardiovascular disease using candidate genes and meta-analysis of association studies. *Ann Hum Genet*. 2006;70:145-169.
4. Morgan TM, Coffey CS, Krumholz HM. Overes-

timation of genetic risks owing to small sample sizes in cardiovascular studies. *Clin Genet*. 2003;64:7-17.

5. Yamada Y. Identification of genetic factors and development of genetic risk diagnosis systems for cardiovascular diseases and stroke. *Circ J*. 2006;70:1240-1248.

6. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet*. 2001;29:306-309.

7. The International HapMap Consortium. The International HapMap Project. *Nature*. 2003;426:789-796.

8. Lutucuta S, Ballantyne CM, Elghannam H, Gotto AM Jr, Marian AJ. Novel polymorphisms in promoter region of ATP binding cassette transporter gene and plasma lipids, severity, progression, and regression of coronary atherosclerosis and response to therapy. *Circ Res*. 2001;88:969-973.

9. Zwarts KY, Clee SM, Zwinderman AH, et al. ABCA1 regulatory variants influence coronary artery disease independent of effects on plasma lipid levels. *Clin Genet*. 2002;61:115-125.

10. Clee SM, Zwinderman AH, Engert JC, et al. Common genetic variation in ABCA1 is associated with altered lipoprotein levels and a modified risk for coronary artery disease. *Circulation*. 2001;103:1198-1205.

11. Tregouet DA, Ricard S, Nicaud V, et al. In-depth haplotype analysis of ABCA1 gene polymorphisms in relation to plasma ApoA1 levels and myocardial infarction. *Arterioscler Thromb Vasc Biol*. 2004;24:775-781.

12. Tobin MD, Braund PS, Burton PR, et al. Genotypes and haplotypes predisposing to myocardial infarction: a multilocus case-control study. *Eur Heart J*. 2004;25:459-467.

13. Zee RY, Cook NR, Reynolds R, Cheng S, Ridker PM. Haplotype analysis of the beta2 adrenergic receptor gene and risk of myocardial infarction in humans. *Genetics*. 2005;169:1583-1587.

14. Higashi K, Ishikawa T, Ito T, Yonemura A, Shige H, Nakamura H. Association of a genetic variation in the beta 3-adrenergic receptor gene with coronary heart disease among Japanese. *Biochem Biophys Res Commun*. 1997;232:728-730.

15. Sethi AA, Nordestgaard BG, Tybjaerg-Hansen A. Angiotensinogen gene polymorphism, plasma angiotensinogen, and risk of hypertension and ischemic heart disease: a meta-analysis. *Arterioscler Thromb Vasc Biol*. 2003;23:1269-1275.

16. Fatini C, Abbate R, Pepe G, et al. Searching for a better assessment of the individual coronary risk profile: the role of angiotensin-converting enzyme, angiotensin II type 1 receptor and angiotensinogen gene polymorphisms. *Eur Heart J*. 2000;21:633-638.

17. Helgadottir A, Manolescu A, Thorleifsson G, et al. The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nat Genet*. 2004;36:233-239.

18. Wang XL, Liu SX, McCredie RM, Wilcken DE. Polymorphisms at the 5′-end of the apolipoprotein AI gene and severity of coronary artery disease. *J Clin Invest*. 1996;98:372-377.

19. Reguero JR, Cubero GI, Batalla A, et al. Apolipoprotein A1 gene polymorphisms and risk of early coronary disease. *Cardiology*. 1998;90:231-235.

20. Wilson PW, Schaefer EJ, Larson MG, Ordovas JM. Apolipoprotein E alleles and risk of coronary disease: a meta-analysis. *Arterioscler Thromb Vasc Biol*. 1996;16:1250-1255.

21. Lambert JC, Brousseau T, Defosse V, et al. Independent association of an APOE gene promoter polymorphism with increased risk of myocardial infarction and decreased APOE plasma concentrations-the ECTIM study. *Hum Mol Genet*. 2000;9:57-61.

22. Aoki S, Mukae S, Itoh S, et al. The genetic factor in acute myocardial infarction with hypertension. *Jpn Circ J*. 2001;65:621-626.

23. Zee RY, Cook NR, Cheng S, et al. Threonine for alanine substitution in the eotaxin (CCL11) gene and the risk of incident myocardial infarction. *Atherosclerosis*. 2004;175:91-94.

24. Ortlepp JR, Vesper K, Mevissen V, et al. Chemokine receptor (CCR2) genotype is associated with myocardial infarction and heart failure in patients under 65 years of age. *J Mol Med*. 2003;81:363-367.

25. González P, Alvarez R, Batalla A, et al. Genetic variation at the chemokine receptors CCR5/CCR2 in myocardial infarction. *Genes Immun*. 2001;2:191-195.

26. Hubacek JA, Rothe G, Pit'ha J, et al. C(−260)→T polymorphism in the promoter of the CD14 monocyte receptor gene as a risk factor for myocardial infarction. *Circulation*. 1999;99:3218-3220.

27. Kuivenhoven JA, Jukema JW, Zwinderman AH, et al; the Regression Growth Evaluation Statin Study Group. The role of a common variant of the cholesteryl ester transfer protein gene in the progression of coronary atherosclerosis. *N Engl J Med*. 1998;338:86-93.

28. Klerkx AH, Tanck MW, Kastelein JJ, et al. Haplotype analysis of the CETP gene: not TaqIB, but the closely linked −629C→A polymorphism and a novel promoter variant are independently associated with CETP concentration. *Hum Mol Genet*. 2003;12:111-123.

29. Eriksson AL, Skrtic S, Niklason A, et al. Association between the low activity genotype of catechol-O-methyltransferase and myocardial infarction in a hypertensive population. *Eur Heart J*. 2004;25:386-391.

30. Niessner A, Marculescu R, Haschemi A, et al. Opposite effects of CX3CR1 receptor polymorphisms V249I and T280M on the development of acute coronary syndrome: a possible implication of fractalkine in inflammatory activation. *Thromb Haemost*. 2005;93:949-954.

31. McDermott DH, Halcox JP, Schenke WH, et al. Association between polymorphism in the chemokine receptor CX3CR1 and coronary vascular endothelial dysfunction and atherosclerosis. *Circ Res*. 2001;89:401-407.

32. Patel S, Steeds R, Channer K, Samani NJ. Analysis of promoter region polymorphism in the aldosterone synthase gene (CYP11B2) as a risk factor for myocardial infarction. *Am J Hypertens*. 2000;13:134-139.

33. Hautanen A, Toivanen P, Manttari M, et al. Joint effects of an aldosterone synthase (CYP11B2) gene polymorphism and classic risk factors on risk of myocardial infarction. *Circulation*. 1999;100:2213-2218.

34. Yasar U, Bennet AM, Eliasson E, et al. Allelic variants of cytochromes P450 2C modify the risk for acute myocardial infarction. *Pharmacogenetics*. 2003;13:715-720.

35. Funk M, Endler G, Freitag R, et al. CYP2C9*2 and CYP2C9*3 alleles confer a lower risk for myocardial infarction. *Clin Chem*. 2004;50:2395-2398.

36. Endler G, Mannhalter C, Sunder-Plassmann H, et al. The K121Q polymorphism in the plasma cell membrane glycoprotein 1 gene predisposes to early myocardial infarction. *J Mol Med*. 2002;80:791-795.

37. Schuit SC, Oei HH, Witteman JC, et al. Estrogen receptor alpha gene polymorphisms and risk of myocardial infarction. *JAMA*. 2004;291:2969-2977.

38. Shearman AM, Cupples LA, Demissie S, et al. Association between estrogen receptor alpha gene variation and cardiovascular disease. *JAMA*. 2003;290:2263-2270.

39. Endler G, Mannhalter C, Sunder-Plassmann H, et al. Homozygosity for the C→T polymorphism at nucleotide 46 in the 5′ untranslated region of the factor XII gene protects from development of acute coronary syndrome. *Br J Haematol*. 2001;115:1007-1009.

40. Endler G, Mannhalter C. Polymorphisms in coagulation factor genes and their impact on arterial and venous thrombosis. *Clin Chim Acta*. 2003;330:31-55.

41. Rosendaal FR, Siscovick DS, Schwartz SM, Psaty BM, Raghunathan TE, Vos HL. A common prothrombin variant (20210 G to A) increases the risk of myocardial infarction in young women. *Blood*. 1997;90:1747-1750.

42. Girelli D, Russo C, Ferraresi P, et al. Polymorphisms in the factor VII gene and the risk of myocardial infarction in patients with coronary artery disease. *N Engl J Med*. 2000;343:774-780.

43. Boekholdt SM, Bijsterveld NR, Moons AH, Levi M, Buller HR, Peters RJ. Genetic variation in coagulation and fibrinolytic proteins and their relation with acute myocardial infarction: a systematic review. *Circulation*. 2001;104:3063-3068.

44. Yamada Y, Izawa H, Ichihara S, et al. Prediction of the risk of myocardial infarction from polymorphisms in candidate genes. *N Engl J Med*. 2002;347:1916-1923.

45. Kenny D, Muckian C, Fitzgerald DJ, Cannon CP, Shields DC. Platelet glycoprotein Ib alpha receptor polymorphisms and recurrent ischaemic events in acute coronary syndrome patients. *J Thromb Thrombolysis*. 2002;13:13-19.

46. Douglas H, Michaelides K, Gorog DA, et al. Platelet membrane glycoprotein Ibalpha gene −5T/C Kozak sequence polymorphism as an independent risk factor for the occurrence of coronary thrombosis. *Heart*. 2002;87:70-74.

47. Lin RC, Wang XL, Morris BJ. Association of coronary artery disease with glucocorticoid receptor N363S variant. *Hypertension*. 2003;41:404-407.

48. Hetet G, Elbaz A, Gariepy J, et al. Association studies between haemochromatosis gene mutations and the risk of cardiovascular diseases. *Eur J Clin Invest*. 2001;31:382-388.

49. Yamada S, Akita H, Kanazawa K, et al. T102C polymorphism of the serotonin (5-HT) 2A receptor gene in patients with non-fatal acute myocardial infarction. *Atherosclerosis*. 2000;150:143-148.

50. Jiang H, Klein RM, Niederacher D, et al. C/T polymorphism of the intercellular adhesion molecule-1 gene (exon 6, codon 469): a risk factor for coronary heart disease and myocardial infarction. *Int J Cardiol*. 2002;84:171-177.

51. Momiyama Y, Hirano R, Taniguchi H, Nakamura H, Ohsuzu F. Effects of interleukin-1 gene polymorphisms on the development of coronary artery disease associated with *Chlamydia pneumoniae* infection. *J Am Coll Cardiol*. 2001;38:712-717.

52. Georges JL, Loukaci V, Poirier O, et al; Etude Cas-Temoin de l'Infarctus du Myocarde. Interleukin-6 gene polymorphisms and susceptibility to myocardial infarction: the ECTIM study. *J Mol Med*. 2001;79:300-305.

53. Jenny NS, Tracy RP, Ogg MS, et al. In the elderly, interleukin-6 plasma levels and the −174G>C polymorphism are associated with the development of cardiovascular disease. *Arterioscler Thromb Vasc Biol*. 2002;22:2066-2071.

54. Baroni MG, D'Andrea MP, Montali A, et al. A common mutation of the insulin receptor substrate-1 gene is a risk factor for coronary artery disease. *Arterioscler Thromb Vasc Biol*. 1999;19:2975-2980.

55. Santoso S, Kunicki TJ, Kroll H, Haberbosch W, Gardemann A. Association of the platelet glycoprotein Ia C807T gene polymorphism with nonfatal myocardial infarction in younger patients. *Blood*. 1999;93:2449-2453.

56. Samara WM, Gurbel PA. The role of platelet receptors and adhesion molecules in coronary artery disease. *Coron Artery Dis*. 2003;14:65-79.

57. Zambon A, Deeb SS, Pauletto P, Crepaldi G, Brunzell JD. Hepatic lipase: a marker for cardiovascular disease risk and response to therapy. *Curr Opin Lipidol*. 2003;14:179-189.

58. Ji J, Herbison CE, Mamotte CD, Burke V, Taylor RR, van Bockxmeer FM. Hepatic lipase gene −514 C/T polymorphism and premature coronary heart disease. *J Cardiovasc Risk*. 2002;9:105-113.

59. Hokanson JE. Functional variants in the lipopro-

tein lipase gene and risk cardiovascular disease. *Curr Opin Lipidol*. 1999;10:393-399.

60. Schulz S, Schagdarsurengin U, Greiser P, et al. The LDL receptor-related protein (LRP1/A2MR) and coronary atherosclerosis–novel genomic variants and functional consequences. *Hum Mutat*. 2002;20:404.

61. PROCARDIS Consortium. A trio family study showing association of the lymphotoxin-alpha N26 (804A) allele with coronary artery disease. *Eur J Hum Genet*. 2004;12:770-774.

62. Ozaki K, Ohnishi Y, Iida A, et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet*. 2002;32:650-654.

63. Herrmann SM, Whatling C, Brand E, et al. Polymorphisms of the human matrix gla protein (MGP) gene, vascular calcification, and myocardial infarction. *Arterioscler Thromb Vasc Biol*. 2000;20:2386-2393.

64. Humphries SE, Martin S, Cooper J, Miller G. Interaction between smoking and the stromelysin-1 (*MMP3*) gene 5A/6A promoter polymorphism and risk of coronary heart disease in healthy men. *Ann Hum Genet*. 2002;66:343-352.

65. Lamblin N, Bauters C, Hermant X, Lablanche JM, Helbecque N, Amouyel P. Polymorphisms in the promoter regions of *MMP-2*, *MMP-3*, *MMP-9* and *MMP-12* genes as determinants of aneurysmal coronary artery disease. *J Am Coll Cardiol*. 2002;40:43-48.

66. Klerk M, Verhoef P, Clarke R, Blom HJ, Kok FJ, Schouten EG. MTHFR 677C→T polymorphism and risk of coronary heart disease: a meta-analysis. *JAMA*. 2002;288:2023-2031.

67. Ledmyr H, McMahon AD, Ehrenborg E, et al. The microsomal triglyceride transfer protein gene-493T variant lowers cholesterol but increases the risk of coronary heart disease. *Circulation*. 2004;109:2279-2284.

68. Juo SH, Han Z, Smith JD, Colangelo L, Liu K. Common polymorphism in promoter of microsomal triglyceride transfer protein gene influences cholesterol, ApoB, and triglyceride levels in young African American men: results from the coronary artery risk development in young adults (CARDIA) study. *Arterioscler Thromb Vasc Biol*. 2000;20:1316-1322.

69. Hyndman ME, Bridge PJ, Warnica JW, Fick G, Parsons HG. Effect of heterozygosity for the methionine synthase 2756 A→G mutation on the risk for recurrent cardiovascular events. *Am J Cardiol*. 2000;86:1144-1146, A1149.

70. Gruchala M, Ciecwierz D, Wasag B, et al. Association of the ScaI atrial natriuretic peptide gene polymorphism with nonfatal myocardial infarction and extent of coronary artery disease. *Am Heart J*. 2003;145:125-131.

71. Tatsuguchi M, Furutani M, Hinagata J, et al. Oxidized LDL receptor gene (*OLR1*) is associated with the risk of myocardial infarction. *Biochem Biophys Res Commun*. 2003;303:247-250.

72. Gardemann A, Mages P, Katz N, Tillmanns H, Haberbosch W. The p22 phox A640G gene polymorphism but not the C242T gene variation is associated with coronary heart disease in younger individuals. *Atherosclerosis*. 1999;145:315-323.

73. Inoue N, Kawashima S, Kanazawa K, Yamada S, Akita H, Yokoyama M. Polymorphism of the NADH/NADPH oxidase p22-phox gene in patients with coronary artery disease. *Circulation*. 1998;97:135-137.

74. Wenzel K, Baumann G, Felix SB. The homozygous combination of Leu125Val and Ser563Asn polymorphisms in the *PECAM1* (*CD31*) gene is associated with early severe coronary heart disease. *Hum Mutat*. 1999;14:545.

75. Andreotti F, Porto I, Crea F, Maseri A. Inflammatory gene polymorphisms and ischaemic heart disease: review of population association studies. *Heart*. 2002;87:107-112.

76. Durrington PN, Mackness B, Mackness MI. Paraoxonase and atherosclerosis. *Arterioscler Thromb Vasc Biol*. 2001;21:473-480.

77. Sanghera DK, Aston CE, Saha N, Kamboh MI. DNA polymorphisms in two paraoxonase genes (*PON1* and *PON2*) are associated with the risk of coronary heart disease. *Am J Hum Genet*. 1998;62:36-44.

78. Ridker PM, Cook NR, Cheng S, et al. Alanine for proline substitution in the peroxisome proliferator-activated receptor gamma-2 (*PPARG2*) gene and the risk of incident myocardial infarction. *Arterioscler Thromb Vasc Biol*. 2003;23:859-863.

79. Cipollone F, Toniato E, Martinotti S, et al. A polymorphism in the cyclooxygenase 2 gene as an inherited protective factor against myocardial infarction and stroke. *JAMA*. 2004;291:2221-2228.

80. Ye L, Miki T, Nakura J, et al. Association of a polymorphic variant of the Werner helicase gene with myocardial infarction in a Japanese population. *Am J Med Genet*. 1997;68:494-498.

81. Herrmann SM, Ricard S, Nicaud V, et al. The P-selectin gene is highly polymorphic: reduced frequency of the Pro715 allele carriers in patients with myocardial infarction. *Hum Mol Genet*. 1998;7:1277-1284.

82. Moatti D, Seknadji P, Galand C, et al. Polymorphisms of the tissue factor pathway inhibitor (TFPI) gene in patients with acute coronary syndromes and in healthy subjects: impact of the V264M substitution on plasma levels of TFPI. *Arterioscler Thromb Vasc Biol*. 1999;19:862-869.

83. Chao TH, Li YH, Chen JH, et al. Relation of thrombomodulin gene polymorphisms to acute myocardial infarction in patients <or =50 years of age. *Am J Cardiol*. 2004;93:204-207.

84. Doggen CJ, Kunz G, Rosendaal FR, et al. A mutation in the thrombomodulin gene, 127G to A coding for Ala25Thr, and the risk of myocardial infarction in men. *Thromb Haemost*. 1998;80:743-748.

85. Wu KK, Aleksic N, Ahn C, Boerwinkle E, Folsom AR, Juneja H. Thrombomodulin Ala455Val polymorphism and risk of coronary heart disease. *Circulation*. 2001;103:1386-1389.

86. Topol EJ, McCarthy J, Gabriel S, et al. Single nucleotide polymorphisms in multiple novel thrombospondin genes may be associated with familial premature myocardial infarction. *Circulation*. 2001;104:2641-2644.

87. Boekholdt SM, Trip MD, Peters RJ, et al. Thrombospondin-2 polymorphism is associated with a reduced risk of premature myocardial infarction. *Arterioscler Thromb Vasc Biol*. 2002;22:e24-e27.

88. Webb KE, Martin JF, Hamsten A, et al. Polymorphisms in the thrombopoietin gene are associated with risk of myocardial infarction at a young age. *Atherosclerosis*. 2001;154:703-711.

89. Kolek MJ, Carlquist JF, Muhlestein JB, et al. Toll-like receptor 4 gene Asp299Gly polymorphism is associated with reductions in vascular inflammation, angiographic coronary artery disease, and clinical diabetes. *Am Heart J*. 2004;148:1034-1040.

90. Padovani JC, Pazin-Filho A, Simoes MV, Marin-Neto JA, Zago MA, Franco RF. Gene polymorphisms in the TNF locus and the risk of myocardial infarction. *Thromb Res*. 2000;100:263-269.

91. Poirier O, Nicaud V, Gariepy J, et al. Polymorphism R92Q of the tumour necrosis factor receptor 1 gene is associated with myocardial infarction and carotid intima-media thickness–the ECTIM, AXA, EVA and GENIC Studies. *Eur J Hum Genet*. 2004;12:213-219.

92. Alpert JS, Thygesen K, Antman E, Bassand JP. Myocardial infarction redefined—a consensus document of the Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction. *J Am Coll Cardiol*. 2000;36:959-969.

93. Braunwald E. Unstable angina: a classification. *Circulation*. 1989;80:410-414.

94. Yan J, Feng J, Hosono S, Sommer SS. Assessment of multiple displacement amplification in molecular epidemiology. *Biotechniques*. 2004;37:136-138, 140-133.

95. Dean FB, Hosono S, Fang L, et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A*. 2002;99:5261-5266.

96. Jurinke C, van den Boom D, Cantor CR, Koster H. The use of MassARRAY technology for high throughput genotyping. *Adv Biochem Eng Biotechnol*. 2002;77:57-74.

97. Jurinke C, Oeth P, van den Boom D. MALDI-TOF mass spectrometry: a versatile tool for high-performance DNA analysis. *Mol Biotechnol*. 2004;26:147-164.

98. Chiodini BD, Barlera S, Franzosi MG, Beceiro VL, Introna M, Tognoni G. APO B gene polymorphisms and coronary artery disease: a meta-analysis. *Atherosclerosis*. 2003;167:355-366.

99. González-Conejero R, Corral J, Roldan V, et al. A common polymorphism in the annexin V Kozak sequence (-1C>T) increases translation efficiency and plasma levels of annexin V, and decreases the risk of myocardial infarction in young patients. *Blood*. 2002;100:2081-2086.

100. Hines LM, Stampfer MJ, Ma J, et al. Genetic variation in alcohol dehydrogenase and the beneficial effect of moderate alcohol consumption on myocardial infarction. *N Engl J Med*. 2001;344:549-555.

101. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*. 2005;76:449-462.

102. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*. 2001;68:978-989.

103. Gauderman WJ. Candidate gene association analysis for a quantitative trait, using parent-offspring trios. *Genet Epidemiol*. 2003;25:327-338.

104. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med*. 2002;21:35-50.

105. Weng L, Kavaslar N, Ustaszewska A, et al. Lack of MEF2A mutations in coronary artery disease. *J Clin Invest*. 2005;115:1016-1020.

106. Liu S, Ma J, Ridker PM, Breslow JL, Stampfer MJ. A prospective study of the association between APOE genotype and the risk of myocardial infarction among apparently healthy men. *Atherosclerosis*. 2003;166:323-329.

107. Freely associating. *Nat Genet*. 1999;22:1-2.

108. Salanti G, Sanderson S, Higgins JP. Obstacles and opportunities in meta-analysis of genetic association studies. *Genet Med*. 2005;7:13-20.

109. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet*. 2004;36:512-517.

110. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science*. 2002;296:2225-2229.

Prediction of Coronary Heart Disease Using Risk Factor Categories

Peter W.F. Wilson, MD; Ralph B. D'Agostino, PhD; Daniel Levy, MD; Albert M. Belanger, BS; Halit Silbershatz, PhD; William B. Kannel, MD

Background—The objective of this study was to examine the association of Joint National Committee (JNC-V) blood pressure and National Cholesterol Education Program (NCEP) cholesterol categories with coronary heart disease (CHD) risk, to incorporate them into coronary prediction algorithms, and to compare the discrimination properties of this approach with other noncategorical prediction functions.

Methods and Results—This work was designed as a prospective, single-center study in the setting of a community-based cohort. The patients were 2489 men and 2856 women 30 to 74 years old at baseline with 12 years of follow-up. During the 12 years of follow-up, a total of 383 men and 227 women developed CHD, which was significantly associated with categories of blood pressure, total cholesterol, LDL cholesterol, and HDL cholesterol (all $P<.001$). Sex-specific prediction equations were formulated to predict CHD risk according to age, diabetes, smoking, JNC-V blood pressure categories, and NCEP total cholesterol and LDL cholesterol categories. The accuracy of this categorical approach was found to be comparable to CHD prediction when the continuous variables themselves were used. After adjustment for other factors, ≈28% of CHD events in men and 29% in women were attributable to blood pressure levels that exceeded high normal (≥130/85). The corresponding multivariable-adjusted attributable risk percent associated with elevated total cholesterol (≥200 mg/dL) was 27% in men and 34% in women.

Conclusions—Recommended guidelines of blood pressure, total cholesterol, and LDL cholesterol effectively predict CHD risk in a middle-aged white population sample. A simple coronary disease prediction algorithm was developed using categorical variables, which allows physicians to predict multivariate CHD risk in patients without overt CHD. (***Circulation*. 1998;97:1837-1847.**)

Key Words: coronary disease ■ prediction ■ hypertension ■ cholesterol

Coronary heart disease continues to be a leading cause of morbidity and mortality among adults in Europe and North America.[1] Risk factors have included blood pressure, cigarette smoking, cholesterol (TC), LDL-C, HDL-C, and diabetes.[2-4] Factors such as obesity, left ventricular hypertrophy, family history of premature CHD, and ERT have also been considered in defining CHD risk.[5-7] Data from population studies enabled prediction of CHD during a follow-up interval of several years, based on blood pressure, smoking history, TC and HDL-C levels, diabetes, and left ventricular hypertrophy on the ECG. These prediction algorithms have been adapted to simplified score sheets that allow physicians to estimate multivariable CHD risk in middle-aged patients.[8]

See p 1761

The present article develops a simplified coronary prediction model, building on the blood pressure, cholesterol, and LDL-C categories proposed by the JNC-V and NCEP ATP II.[7,9,10] The analysis evaluates the utility and accuracy of blood pressure, cholesterol, and LDL-C recommended categories in multivariable CHD prediction, using a Framingham Heart Study sample that pooled information for the original and offspring cohorts and followed them for 12 years. This approach emphasizes the established, powerful, independent, and biologically important factors. Family history for heart disease, physical activity, and obesity are not included because these factors work to a large extent through the major risk factors, and their unique contribution to CHD prediction can be difficult to quantify. The prediction of initial CHD events in a free-living population not on medication is emphasized. Consequently, ERT for postmenopausal women, treatment of high blood pressure, and therapy for high blood cholesterol are not included in the formulations.

Methods

The population-based sample used for this report included 2489 men and 2856 women 30 to 74 years old at the time of their Framingham Heart Study examination in 1971 to 1974. Participants attended either the 11th examination of the original Framingham cohort[11] or the initial examination of the Framingham Offspring Study.[12] Similar research protocols were used in each study, and persons with overt CHD at the baseline examination were excluded.

From the Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, Mass (P.W.F.W., D.L.); Boston University Mathematics Department, Boston, Mass (R.B.D., A.M.B., H.S.); and Framingham Heart Study, Boston University School of Medicine, Framingham, Mass (W.B.K.).

Reprint requests to Dr Peter W.F. Wilson, Framingham Heart Study, National Heart, Lung, and Blood Institute, 5 Thurber St, Framingham, MA 01701. E-mail peter@fram.nhlbi.nih.gov Score sheets are on the internet at http://www.nhlbi.nih.gov/nhlbi/fram/

Selected Abbreviations and Acronyms

CHD	=	coronary heart disease
ERT	=	estrogen replacement therapy
HDL-C	=	HDL cholesterol
JNC-V	=	Fifth Joint National Committee on Hypertension
LDL-C	=	LDL cholesterol
NCEP ATP II	=	National Cholesterol Education Program, Adult Treatment Panel II
TC	=	total cholesterol
VLDL-C	=	VLDL cholesterol

At the 1971–1974 examination, a medical history was taken and a physical examination was performed by a physician. Persons who smoked regularly during the previous 12 months were classified as smokers. Height and weight were measured, and body mass index (kg/m^2) was calculated. Two blood pressure determinations were made after the participant had been sitting at least 5 minutes, and the average was used for analyses. Hypertension was categorized according to blood pressure readings by JNC-V definitions[10]: optimal (systolic <120 mm Hg and diastolic <80 mm Hg), normal blood pressure (systolic 120 to 129 mm Hg or diastolic 80 to 84 mm Hg), high normal blood pressure (systolic 130 to 139 mm Hg or diastolic 85 to 89 mm Hg), hypertension stage I (systolic 140 to 159 mm Hg or diastolic 90 to 99 mm Hg), and hypertension stage II–IV (systolic ≥160 or diastolic ≥100 mm Hg). When systolic and diastolic pressures fell into different categories, the higher category was selected for the purposes of classification. Blood pressure categorization was made without regard to the use of antihypertensive medication.

Diabetes was considered present if the participant was under treatment with insulin or oral hypoglycemic agents, if casual blood glucose determinations exceeded 150 mg/dL at two clinic visits in the original cohort, or if fasting blood glucose exceeded 140 mg/dL at the initial examination of the Offspring Study participants. Blood was drawn at the baseline examination after an overnight fast, and EDTA plasma was used for all cholesterol and triglyceride measurements. Cholesterol was determined according to the Abell-Kendall technique,[13] and HDL-C was measured after precipitation of VLDL and LDL proteins with heparin-magnesium according to the Lipid Research Clinics Program protocol.[14] When triglycerides were <400 mg/dL, the concentration of LDL-C was estimated indirectly by use of the Friedewald formula[15]; for triglycerides ≥400 mg/dL, the LDL-C was estimated directly after ultracentrifugation of plasma and measurement of cholesterol in the bottom fraction (plasma density <1.006).[16]

Cutoffs for TC (<200, 200 to 239, 240 to 279, and ≥280 mg/dL), LDL-C (<130, 130 to 159, and ≥160 mg/dL), HDL-C (<35, 35 to 59, and ≥60 mg/dL), cigarette smoking, diabetes, and age were considered in this report. The cholesterol and LDL-C cutoffs are similar to those used for the NCEP ATP II guidelines and were partly dictated by the number of persons with higher levels of TC or LDL-C. For those reasons, we have provided information for cholesterol categories of 240 to 279 and ≥280 mg/dL and for LDL-C ≥160 mg/dL. Too few persons had LDL-C ≥190 mg/dL to provide stable estimates for CHD risk. Study subjects were followed up over a 12-year period for the development of CHD (angina pectoris, recognized and unrecognized myocardial infarction, coronary insufficiency, and coronary heart disease death) according to previously published criteria. "Hard CHD" events included total CHD without angina pectoris.[17] Surveillance for CHD consisted of regular examinations at the Framingham Heart Study clinic and review of medical records from outside physician office visits and hospitalizations.

Statistical tests included age-adjusted linear regression or logistic regression to test for trends across blood pressure, TC, LDL-C, and HDL-C categories.[18] Age-adjusted Cox proportional hazards regression and its accompanying c statistic were used to test for the relation between various independent variables and the CHD outcome and to evaluate the discriminatory ability of various prediction models.[19,20] The 12-year follow-up was used in the proportional hazards models, and results were adapted to provide 10-year CHD incidence estimates. Separate score sheets were developed for each sex using TC and LDL-C categories. These sheets adapted the results of proportional hazards regressions by use of a system that assigned points for each risk factor based on the value for the corresponding β-coefficient of the regression analyses.

The relative risk, but not the attributable risk, for TC and CHD declines with advancing age.[21] Quadratic terms for age were considered in the models for the score sheets. Furthermore, CHD risk is associated with HDL-C in the elderly,[22-24] and interaction terms for TC and age were also considered in the development of the prediction models.[22] Among women, an age-squared term was found to be significant in the prediction models and was incorporated into the score sheets. Neither age×TC nor age×LDL-C was found to be significant in either sex.

Score sheets for prediction of CHD using TC and LDL-C categorical variables were developed from the β-coefficients of Cox proportional hazards models. The TC range was expanded in 40-mg/dL increments to include ≥160 mg/dL and ≥280 mg/dL, the HDL-C range 35 to 59 mg/dL was partitioned to provide three levels for each sex, and both optimal and normal blood pressure categories were included. The score sheets provide comparison 10-year absolute risks for persons of the same age and sex for average total CHD, average hard CHD (total CHD without angina pectoris), and low-risk total CHD. Risk factors are shaded, ranging from very low relative risk to very high. Such distinctions are arbitrary but provide a foundation to determine the need for clinical intervention.

Results

At initial examination, study subjects ranged in age from 30 to 74 years, and the mean age±SD was 48.6±11.7 years for 2489 men and 49.8±12.0 years for 2856 women. Because there were relatively few persons at the higher stages of hypertension in the Framingham sample, stages II, III, and IV hypertension were combined into a single category in the analyses (Table 1). Approximately half of the subjects for each sex had blood pressure levels in the normal or optimal range.

The age-adjusted means for various risk factors according to blood pressure categories are shown for men and women in Table 2. Therapy for hypertension (*P*<.001 men, *P*<.001 women), more frequent diabetes (*P*<.001 men, *P*<.001 women), greater body

TABLE 1. Characteristics of Participants According to JNC-V Hypertension Categories*

	Blood Pressure			
	Systolic, mm Hg	Diastolic, mm Hg	Men, %	Women, %
Normal (including optimal)	<130	<85	44	55
High normal	130–139	85–89	20	15
Hypertension stage I	140–159	90–99	23	19
Hypertension stage II–IV	≥160	≥100	13	11

*Ignoring blood pressure therapy.

TABLE 2. Age-Adjusted Mean Levels and Prevalence of Risk Factors According to Blood Pressure Category

	Not Hypertensive		Hypertensive		
	Normal	High Normal	Stage I	Stage II–IV	*P*, Test for Trend*
Men	(n=1097)	(n=500)	(n=567)	(n=325)	
Hypertensive therapy, %	1.6	2.7	10.1	25.0	<.001
Body mass index, kg/m^2	25.8	26.7	27.5	28.3	<.001
Cigarette use, %	43.1	41.8	35.4	38.2	.010
Diabetes, %	3.6	6.1	4.0	11.2	<.001
TC, mg/dL	210.1	214.3	218.0	213.9	.004
LDL-C, mg/dL	142.7	143.4	144.5	139.7	.638
HDL-C, mg/dL	44.4	45.7	44.8	44.5	.674
Women	(n=1578)	(n=424)	(n=535)	(n=319)	
Hypertensive therapy, %	3.9	9.4	18.0	33.6	<.001
Body mass index, kg/m^2	23.9	25.8	26.3	26.9	<.001
Cigarette use, %	39.4	37.3	33.9	35.9	.071
Diabetes, %	2.6	3.4	4.9	9.8	<.001
TC, mg/dL	214.1	223.0	224.4	218.5	<.001
LDL-C, mg/dL	138.3	143.9	146.8	138.9	.031
HDL-C, mg/dL	58.6	58.2	55.9	55.7	<.001

*Test for linear trend across blood pressure categories after age adjustment. For dichotomous variables, logistic regression was done.

mass index ($P<.001$ men, $P<.001$ women), and higher TC level ($P=.004$ men, $P<.001$ women) were consistently associated with higher blood pressure categories in both sexes. Cigarette smoking was inversely associated with blood pressure in men ($P=.010$), but only a borderline association was present in women ($P=.071$). The lipoprotein fractions HDL-C ($P<.001$) and LDL-C ($P=.031$) were significantly associated with blood pressure category in women but not in men.

Age-adjusted 10-year CHD rates for blood pressure and cholesterol categories are shown for men and women in Table 3. In prediction models, the CHD rates were significantly associated with the specified categories of blood pressure, TC, HDL-C, and LDL-C (all $P<.001$ for both sexes). The number of CHD events arising at each blood pressure and cholesterol category is also given. For blood pressure, the greatest number of CHD cases arose from the stage I hypertension category for both sexes. Conversely, the greatest number of CHD cases arose from the highest lipoprotein cholesterol levels (LDL-C ≥160 mg/dL or cholesterol ≥240 mg/dL).

Multivariable risk calculations for TC categories are shown in Table 4. Normal or optimal blood pressure was used as the reference level, and estimated relative risk rose from 1.00 for normal or optimal blood pressure to 1.84 in men and 2.12 in women with stage II–IV hypertension. Similarly, for TC, the estimated relative risk rose from 1.00 for levels <200 mg/dL to 1.90 in men and 1.72 in women with TC ≥240 mg/dL. When typical HDL-C levels (35 to 59 mg/dL) were used as a reference, CHD risk was increased among men and women with low HDL-C (<35 mg/dL) and CHD risk was correspondingly decreased among subjects with high HDL-C (≥60 mg/dL). The population-attributable risk percent associated with hypertension was 6% for high normal, 13% for stage I, and 9% for stage II–IV hypertension among men. The corresponding values were 5% for high normal, 13% for stage I, and 12% for stage II–IV hypertension among women. An overall estimate of the attributable risk percent for blood pressure level greater than normal was 28% in men and 29% in women. When cholesterol <200 mg/dL was used as the reference range, attributable risks were 10% for TC 200 to 239 mg/dL and 17% for TC ≥240 mg/dL in men and 12% for TC 200 to 239 mg/dL and 22% for TC ≥240 mg/dL in women. The overall estimate of the attributable risk percent for TC level ≥200 mg/dL was 27% in men and 34% in women.

Multivariable risk calculations for LDL-C categories are shown in Table 5, and these results parallel the presentation in Table 4. When LDL-C <130 mg/dL is used as the reference range, a greater absolute CHD risk is associated with higher LDL-C categories, but the magnitude of the relative risk and its statistical significance are very similar to that observed for the categories of TC (Table 4).

The efficacy of prediction with continuous variables was compared with that obtained with categorical variables and a risk factor sum (Figs 1 and 2 for men and women, respectively). For calculation of the risk factor sum, the levels considered were age (≥45 years for men, ≥55 years for women), hypertension (systolic blood pressure ≥140 mm Hg, diastolic blood pressure ≥90 mm Hg, or use of antihypertensive medication), smoking, diabetes, elevated cholesterol (cholesterol ≥240 mg/dL or LDL-C ≥160 mg/dL), and HDL-C <35 mg/dL. One point was given for each risk factor, for a possible score of 0 to 7 points. A greater area under the curve indicated better predictive capability. The curves were nearly identical for the continuous and categorical formulations, TC and LDL-C categories had similar effects, and the risk factor sums tended to have the lowest predictive potential. The c statistic, a measure of the discriminatory ability of a model, equal to the area under the receiver operating characteristic curve, provides a guide to interpret the

TABLE 3. CHD Risk According to Blood Pressure and Lipid Categories

	Men			Women		
	Person-Years	No. of Events (%)	Age-Adjusted 10-Year Rate	Person-Years	No. of Events (%)	Age-Adjusted 10-Year Rate
Total	30 154	383 (100)		38 057	227 (100)	
Blood pressure						
Normal (including optimal)	13 524	110 (29)	7.8	20 747	66 (29)	2.9
High normal	6307	77 (20)	12.4	6056	36 (16)	7.1
Hypertension stage I	6695	115 (30)	16.0	7254	72 (32)	13.9
Hypertension stage II–IV	3628	81 (21)	20.9	4000	53 (23)	14.1
TC, mg/dL						
<200	11 591	103 (27)	8.2	13 289	39 (17)	3.1
200–239	11 792	148 (39)	12.0	12 683	80 (35)	6.6
≥240	6771	132 (34)	18.6	12 085	108 (48)	10.3
HDL-C, mg/dL						
<35	5601	97 (25)	15.8	1506	23 (10)	14.7
35–59	21 151	260 (68)	12.0	20 788	146 (64)	7.5
≥60	3409	26 (7)	8.2	15 761	58 (26)	3.9
LDL-C, mg/dL						
<130	11 142	104 (27)	7.3	15 835	50 (22)	2.3
130–159	10 384	124 (32)	11.3	10 455	64 (28)	6.5
≥160	8628	155 (41)	17.3	11 767	113 (50)	10.6

The age-adjusted 10-year CHD rates were calculated from the Cox proportional hazards model, based on 12 years of follow-up.

results plotted in Figs 1 and 2. The c statistics associated with TC categories were 0.74 in men and 0.77 in women for continuous variables by proportional hazards or accelerated failure models,[11] 0.73 in men and 0.76 in women for categorical variables, and 0.69 in men and 0.72 in women for the risk factor sum. The corresponding c statistics associated with LDL-C categories were 0.74 in men and 0.77 in women for continuous variables by proportional hazards or accelerated failure models,[11] 0.73 in men and 0.77 in women for categorical variables, and 0.68 in men and 0.71 in women for the risk factor sum.

TABLE 4. Multivariable-Adjusted Relative Risks for CHD According to TC Categories

	Men		Women	
	Relative Risk	95% CI	Relative Risk	95% CI
Age, y	1.05‡	1.04–1.06	1.04‡	1.03–1.06
Blood pressure				
Normal (including optimal)	1.00	Referent	1.00	Referent
High normal	1.31	0.98–1.76	1.30	0.86–1.98
Hypertension stage I	1.67†	1.28–2.18	1.73†	1.19–2.52
Hypertension stage II–IV	1.84‡	1.37–2.49	2.12†	1.42–3.17
Cigarette use (y/n)	1.68‡	1.37–2.06	1.47†	1.12–1.94
Diabetes (y/n)	1.50*	1.06–2.13	1.77†	1.16–2.69
TC, mg/dL				
<200	1.00	Referent	1.00	Referent
200–239	1.31*	1.01–1.68	1.51*	1.01–2.24
≥240	1.90‡	1.47–2.47	1.72†	1.15–2.56
HDL-C, mg/dL				
<35	1.47†	1.16–1.86	2.02†	1.29–3.15
35–59	1.00	Referent	1.00	Referent
≥60	0.56†	0.37–0.83	0.58†	0.43–0.79

The multivariate models were performed separately for men and women. Each model included simultaneously all variables listed in the table. All analyses used categorical variables.

*.01<*P*<.05, †.001<*P*<.01, ‡*P*<.001.

TABLE 5. Multivariate-Adjusted Relative Risks for CHD According to LDL-C Categories

	Men		Women	
	Relative Risk	95% CI	Relative Risk	95% CI
Age, y	1.05‡	1.04–1.06	1.04‡	1.03–1.06
Blood pressure				
Normal (including optimal)	1.00	Referent	1.00	Referent
High normal	1.32	0.98–1.78	1.34	0.88–2.05
Hypertension stage I	1.73‡	1.32–2.26	1.75†	1.21–2.54
Hypertension stage II	1.92‡	1.42–2.59	2.19‡	1.46–3.27
Cigarette use (y/n)	1.71‡	1.39–2.10	1.49†	1.13–1.97
Diabetes (y/n)	1.47*	1.04–2.08	1.80†	1.18–2.74
LDL-C, mg/dL				
<130	1.00	Referent	1.00	Referent
130–159	1.19	0.91–1.54	1.24	0.84–1.81
≥160	1.74‡	1.36–2.24	1.68†	1.17–2.40
HDL-C, mg/dL				
<35	1.46†	1.15–1.85	2.08†	1.33–3.25
35–59	1.00	Referent	1.00	Referent
≥60	0.61*	0.41–0.91	0.64†	0.47–0.87

The multivariate models were performed separately for men and women. Each model included simultaneously all variables listed in the table. All analyses used categorical variables.

*.01<P<.05, †.001<P<.01, ‡P<.001.

Score sheets were developed to predict CHD in men (Fig 3) and women (Fig 4) from the β-coefficients of Cox proportional hazards models (Table 6). Among women, an age-squared term was found to be significant and was incorporated into the score sheets. The average CHD risk over a period of 10 years tends to plateau slightly in the oldest men and women.

An illustrative example for Fig 3 follows. The subject is a 55-year-old man with a TC of 250 mg/dL, HDL-C of 39 mg/dL, and blood pressure of 146/88 who is diabetic and a nonsmoker. Proceeding through the steps gives us the following results: Step 1: Age 55=4 points. Step 2: TC 250 mg/dL=2 points. Step 3: HDL-C 39 mg/dL=1 point. Step 4: Blood pressure 146/88 mm Hg=2 points. Step 5: Diabetic=2 points. Step 6: Nonsmoker=0 points. Step 7: Point total was 4+2+1+2+2+0=11. Step 8: Estimated 10-year CHD risk is 31%. Step 9: The average and "low-risk" risks of CHD over a period of 10 years for a 55-year-old man are 16% and 7%, respectively (low risk was calculated for a person the same age, optimal blood pressure, TC 160 to 199 mg/dL, HDL-C 45 mg/dL for men or 55 mg/dL for women, nonsmoker, and no diabetes). Dividing the subject's risk by the

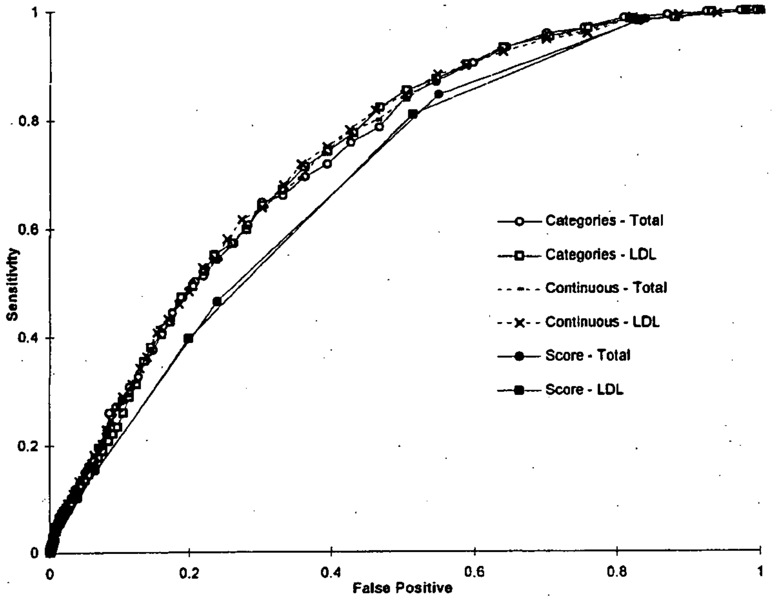


Figure 1. Receiver operating characteristic curves for prediction of CHD in Framingham men over a period of 12 years. Separate plots were used for continuous, categorical, and risk factor sum models, according to whether TC or calculated LDL-C was used.

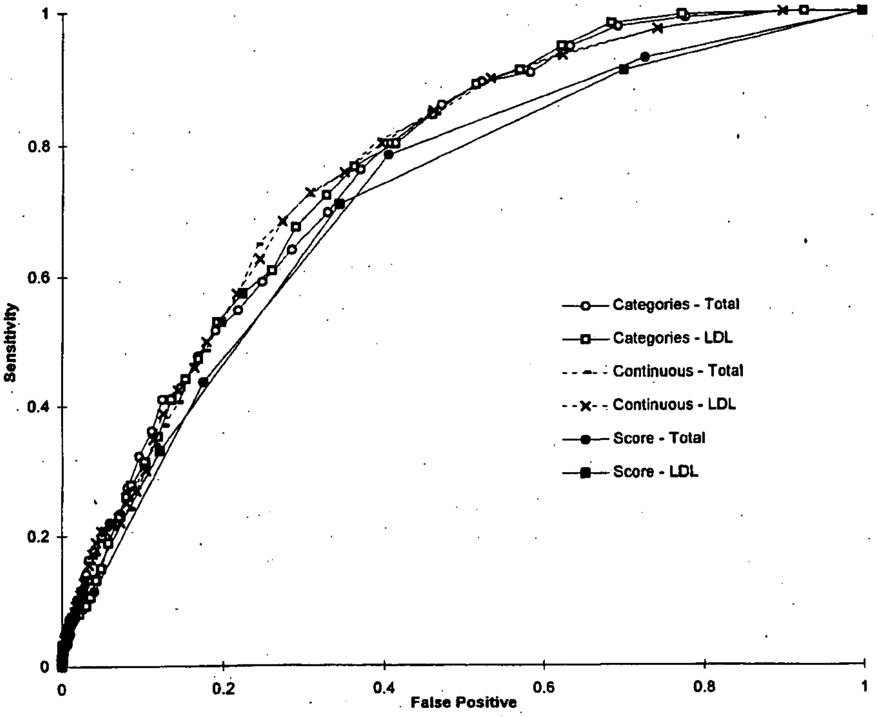


Figure 2. Receiver operating characteristic curves for prediction of CHD in Framingham women over a period of 12 years. Separate plots were used for continuous, categorical, and risk factor sum models, according to whether TC or calculated LDL-C were used.

average risk provides an estimate of the relative risk: 31% divided by 16%=1.94. Use of the LDL-C approach in the score sheets is appropriate when fasting LDL-C estimates are available, by use of ultracentrifugation techniques, the Friedewald formula, or newer LDL-C assays.[15,25,26] The approach is analogous to that shown for TC categories.

Discussion

For the past two decades it has been possible to estimate CHD risk by use of regression equations derived from observational studies, and the present study demonstrates similar results, predicting later CHD in a middle-aged white population sample. Prediction models have typically been based on the logistic function, although the Weibull distribution has also been used.[11,22] Formulations have often included age, sex, blood pressure, TC, HDL-C, smoking, diabetes, and left ventricular hypertrophy.[11] The prediction of CHD has taken the form of sex-specific equations that were developed from a single study and applied to other populations or individuals. Age, TC, HDL-C, and blood pressure were used in the equations as continuous variables, in contrast to dichotomous variables (yes/no) such as smoking, diabetes, and left ventricular hypertrophy.

The present study builds on the prior experience of CHD prediction with continuous variables and integrates the categorical approaches that have become part of the framework of blood pressure (JNC-V) and cholesterol (NCEP) programs in the United States.[6,7,10] As suggested in an earlier NCEP report,[27] our approach integrates blood pressure and cholesterol information and estimates both relative and absolute CHD risk with a risk factor weighting approach.

The NCEP ATP II guidelines defined hypertension as a yes/no variable, and it can be seen from Tables 3, 4, and 5 that additional blood pressure categories are important in predicting CHD risk. Higher levels of blood pressure are typically associated with abnormal cholesterol levels, greater body mass index, and an increased prevalence of diabetes (Table 2). Data from Tables 3 and 4 demonstrate that blood pressure, TC, LDL-C, and HDL-C categories are predictive of CHD and suggest that risk factor prevention and intervention programs should be integrated, as recently suggested.[28–30] Three reasons probably account for similar results when continuous or categorical formulations are used: (1) a large enough number of categories has been used to adequately describe the clinical data; (2) coronary prediction equations have limitations in their precision and accuracy; and (3) in the final steps of the prediction score sheet, the data are summarized, by use of point score totals, providing fewer than 20 combinations for CHD risk prediction.

The predictive capability of the continuous model described here is similar to the accelerated failure model used in an earlier Framingham CHD prediction equation,[11] and the continuous variable and categorical variable approaches have c-statistic values that are nearly identical, suggesting that predictability of the models is nearly the same in either instance. This result is in contradistinction to a comparison of the NCEP ATP II algorithm (<10 unique patterns) with a continuous variable approach in which the latter (using Framingham models) was thought to be statistically superior.[29] A risk factor sum model, considering 7 dichotomous variables, was used for comparison in the present study and showed a significant falloff in the level of the c statistic with this approach compared with formulations using categorical or continuous levels.

TC- and LDL-C–based approaches, whether continuous or categorical variables are used, are similar in their ability to predict initial CHD events in the models presented. This may result from indirect estimation of LDL-C, leading to reduced

Step 1

Age		
Years	LDL Pts	Chol Pts
30-34	-1	[-1]
35-39	0	[0]
40-44	1	[1]
45-49	2	[2]
50-54	3	[3]
55-59	4	[4]
60-64	5	[5]
65-69	6	[6]
70-74	7	[7]

Step 2

LDL - C		
(mg/dl)	(mmol/L)	LDL Pts
<100	<2.59	-3
100-129	2.60-3.36	0
130-159	3.37-4.14	0
160-190	4.15-4.92	1
≥190	≥4.92	2

Cholesterol		
(mg/dl)	(mmol/L)	Chol Pts
<160	<4.14	[-3]
160-199	4.15-5.17	[0]
200-239	5.18-6.21	[1]
240-279	6.22-7.24	[2]
≥280	≥7.25	[3]

Step 3

HDL - C			
(mg/dl)	(mmol/L)	LDL Pts	Chol Pts
<35	<0.90	2	[2]
35-44	0.91-1.16	1	[1]
45-49	1.17-1.29	0	[0]
50-59	1.30-1.55	0	[0]
≥60	≥1.56	-1	[-2]

Step 4

Blood Pressure					
Systolic	Diastolic (mm Hg)				
(mm Hg)	<80	80-84	85-89	90-99	≥100
<120	0 [0] pts				
120-129		0 [0] pts			
130-139			1 [1] pts		
140-159				2 [2] pts	
≥160					3 [3] pts

Note: When systolic and diastolic pressures provide different estimates for point scores, use the higher number

Step 5

Diabetes		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

Step 6

Smoker		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

(sum from steps 1-6)

Step 7

Adding up the points

Age ______

LDL-C or Chol ______

HDL - C ______

Blood Pressure ______

Diabetes ______

Smoker ______

Point total ______

(determine CHD risk from point total)

Step 8

CHD Risk			
LDL Pts Total	10 Yr CHD Risk	Chol Pts Total	10 Yr CHD Risk
≤-3	1%		
-2	2%		
-1	2%	[≤-1]	[2%]
0	3%	[0]	[3%]
1	4%	[1]	[3%]
2	4%	[2]	[4%]
3	6%	[3]	[5%]
4	7%	[4]	[7%]
5	9%	[5]	[8%]
6	11%	[6]	[10%]
7	14%	[7]	[13%]
8	18%	[8]	[16%]
9	22%	[9]	[20%]
10	27%	[10]	[25%]
11	33%	[11]	[31%]
12	40%	[12]	[37%]
13	47%	[13]	[45%]
≥14	≥56%	[≥14]	[≥53%]

(compare to average person your age)

Step 9

Comparative Risk			
Age (years)	Average 10 Yr CHD Risk	Average 10 Yr Hard* CHD Risk	Low** 10 Yr CHD Risk
30-34	3%	1%	2%
35-39	5%	4%	3%
40-44	7%	4%	4%
45-49	11%	8%	4%
50-54	14%	10%	6%
55-59	16%	13%	7%
60-64	21%	20%	9%
65-69	25%	22%	11%
70-74	30%	25%	14%

Key	
Color	Relative Risk
green	Very low
white	Low
yellow	Moderate
rose	High
red	Very high

* Hard CHD events exclude angina pectoris

** Low risk was calculated for a person the same age, optimal blood pressure, LDL-C 100-129 mg/dL or cholesterol 160-199 mg/dl, HDL-C 45 mg/dL for men or 55 mg/dL for women, non-smoker, no diabetes

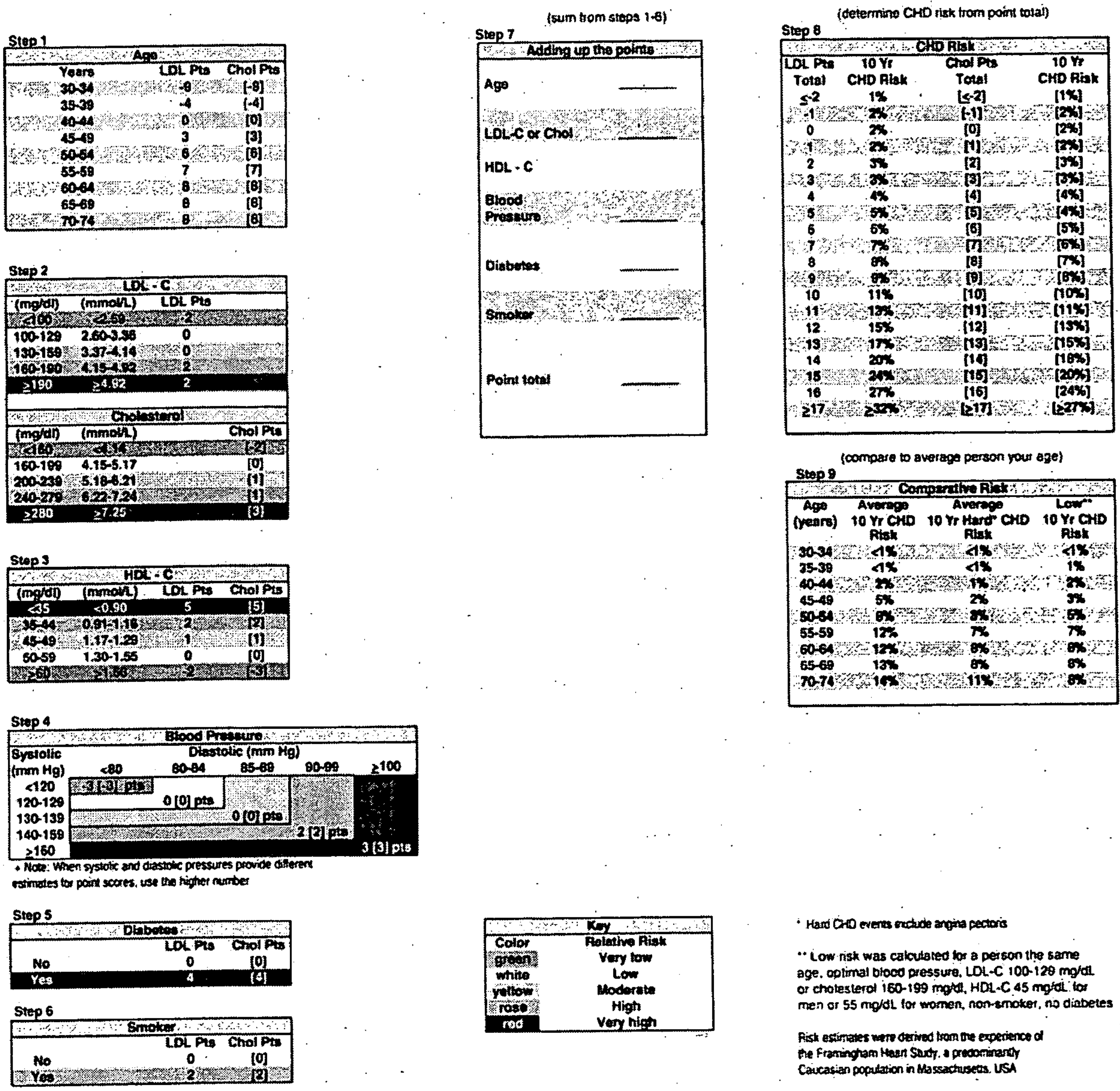
Risk estimates were derived from the experience of the Framingham Heart Study, a predominantly Caucasian population in Massachusetts, USA

Figure 3. CHD score sheet for men using TC or LDL-C categories. Uses age, TC (or LDL-C), HDL-C, blood pressure, diabetes, and smoking. Estimates risk for CHD over a period of 10 years based on Framingham experience in men 30 to 74 years old at baseline. Average risk estimates are based on typical Framingham subjects, and estimates of idealized risk are based on optimal blood pressure, TC 160 to 199 mg/dL (or LDL 100 to 129 mg/dL), HDL-C of 45 mg/dL in men, no diabetes, and no smoking. Use of the LDL-C categories is appropriate when fasting LDL-C measurements are available. Pts indicates points.

accuracy and precision of LDL-C estimates from single blood measurements.[31,32] The CHD estimates in the present article represent the experience of a free-living population sample, and different results may be obtained when blood pressure or blood cholesterol has been treated aggressively.

Although the impact of TC and LDL-C on estimates of CHD risk is similar in Framingham data, such results may be more relevant to populations than to individuals. Extensive clinical data and clinical trial results suggest that LDL-C is the major atherogenic lipoprotein and that measurement of LDL-C levels in the clinical setting provides an advantage.[33-35] High or low levels of HDL-C within individuals can produce discrepancies between TC and LDL-C levels. In addition, TC and LDL-C levels are not always concordant in persons with hypertriglyceridemia. Thus, measurement of TC is only a crude surrogate for LDL-C in risk assessment or in estimating initial response to therapy, although it can be useful in initial detection or long-term monitoring of response.[31]

Several candidate variables were not used in the prediction equations. A family history of premature CHD, previously shown in the Framingham Study to increase the relative odds of CHD to ≈1.3,[36] was not uniformly



Step 1

Age		
Years	LDL Pts	Chol Pts
30-34	-9	[-9]
35-39	-4	[-4]
40-44	0	[0]
45-49	3	[3]
50-54	6	[6]
55-59	7	[7]
60-64	8	[8]
65-69	8	[8]
70-74	8	[8]

Step 2

LDL - C		
(mg/dl)	(mmol/L)	LDL Pts
<100	<2.59	-2
100-129	2.60-3.36	0
130-159	3.37-4.14	0
160-190	4.15-4.92	2
≥190	≥4.92	2

Cholesterol		
(mg/dl)	(mmol/L)	Chol Pts
<160	<4.14	[-2]
160-199	4.15-5.17	[0]
200-239	5.18-6.21	[1]
240-279	6.22-7.24	[1]
≥280	≥7.25	[3]

Step 3

HDL - C			
(mg/dl)	(mmol/L)	LDL Pts	Chol Pts
<35	<0.90	5	[5]
35-44	0.91-1.16	2	[2]
45-49	1.17-1.29	1	[1]
50-59	1.30-1.55	0	[0]
≥60	≥1.56	-2	[-3]

Step 4

Blood Pressure					
Systolic	Diastolic (mm Hg)				
(mm Hg)	<80	80-84	85-89	90-99	≥100
<120	-3 [-3] pts				
120-129		0 [0] pts			
130-139			0 [0] pts		
140-159				2 [2] pts	
≥160					3 [3] pts

• Note: When systolic and diastolic pressures provide different estimates for point scores, use the higher number

Step 5

Diabetes		
	LDL Pts	Chol Pts
No	0	[0]
Yes	4	[4]

Step 6

Smoker		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

(sum from steps 1-6)

Step 7

Adding up the points

Age ______

LDL-C or Chol ______

HDL - C ______

Blood Pressure ______

Diabetes ______

Smoker ______

Point total ______

Key	
Color	Relative Risk
green	Very low
white	Low
yellow	Moderate
rose	High
red	Very high

(determine CHD risk from point total)

Step 8

CHD Risk			
LDL Pts Total	10 Yr CHD Risk	Chol Pts Total	10 Yr CHD Risk
≤-2	1%	[≤-2]	[1%]
-1	2%	[-1]	[2%]
0	2%	[0]	[2%]
1	2%	[1]	[2%]
2	3%	[2]	[3%]
3	3%	[3]	[3%]
4	4%	[4]	[4%]
5	5%	[5]	[4%]
6	6%	[6]	[5%]
7	7%	[7]	[6%]
8	8%	[8]	[7%]
9	9%	[9]	[8%]
10	11%	[10]	[10%]
11	13%	[11]	[11%]
12	15%	[12]	[13%]
13	17%	[13]	[15%]
14	20%	[14]	[18%]
15	24%	[15]	[20%]
16	27%	[16]	[24%]
≥17	≥32%	[≥17]	[≥27%]

(compare to average person your age)

Step 9

Comparative Risk			
Age (years)	Average 10 Yr CHD Risk	Average 10 Yr Hard* CHD Risk	Low** 10 Yr CHD Risk
30-34	<1%	<1%	<1%
35-39	<1%	<1%	1%
40-44	2%	1%	2%
45-49	5%	2%	3%
50-54	8%	3%	5%
55-59	12%	7%	7%
60-64	12%	8%	8%
65-69	13%	8%	8%
70-74	14%	11%	8%

* Hard CHD events exclude angina pectoris

** Low risk was calculated for a person the same age, optimal blood pressure, LDL-C 100-129 mg/dL or cholesterol 160-199 mg/dl, HDL-C 45 mg/dL for men or 55 mg/dL for women, non-smoker, no diabetes

Risk estimates were derived from the experience of the Framingham Heart Study, a predominantly Caucasian population in Massachusetts, USA

Figure 4. CHD score sheet for women using TC or LDL-C categories. Uses age, TC, HDL-C, blood pressure, diabetes, and smoking. Estimates risk for CHD over a period of 10 years based on Framingham experience in women 30 to 74 years old at baseline. Average risk estimates are based on typical Framingham subjects, and estimates of idealized risk are based on optimal blood pressure, TC 160 to 199 mg/dL (or LDL 100 to 129 mg/dL), HDL-C of 55 mg/dL in women, no diabetes, and no smoking. Use of the LDL-C categories is appropriate when fasting LDL-C measurements are available. Pts indicates points.

available among the second-generation participants. Fibrinogen is now recognized as a CHD risk factor,[37] and levels were available for ≈1000 original cohort participants at a 1968–70 examination,[38,39] but fibrinogen measurements were not available for the Offspring Study participants. In addition, established methods for measuring fibrinogen are lacking, and the precise mechanism linking elevated fibrinogen levels to CHD is unclear. Other risk factors, such as smoking, diabetes, and hypertension, are often associated with abnormal fibrinogen levels, and fibrinogen measurements vary greatly within individuals.[37,40] Left ventricular hypertrophy on the ECG was used in previous CHD prediction algorithms, but it is highly associated with hypertension and was not included in the present formulation for a variety of reasons, including lack of standard universally accepted ECG criteria.[11]

Postmenopausal ERT was not used in the prediction algorithm, because estrogen dose was typically higher in the early 1970s[41] and the cardioprotective effects of hormonal replacement therapy that have been universally observed in more recent times[42–45] were not experienced by all Framingham women from the early 1970s to the mid 1980s.[46–48]

Persons who exercise typically have a lower risk of CHD.[49–51] Information on physical activity was not available at the baseline examinations used to develop this CHD risk prediction algorithm, but cigarette smoking, low HDL-C levels, and diabetes are less common among those who are physically active.[52–55] Regular and vigorous exercise is often associated with higher levels of HDL-C, an important determinant for reduced CHD risk.[56–58] Similarly, body mass index, an obesity index that expresses weight in kilograms divided by height in meters squared, has been considered a candidate variable for the CHD prediction algorithm. Greater obesity has been associated with higher TC, lower HDL-C, higher blood pressure, and diabetes, and the residual impact of obesity on CHD has typically been slight after incorporation of these other variables into the regression model.[8]

Clinicians should exercise caution in generalizing from experience of the Framingham Study, a community sample of white subjects drawn from a suburb west of Boston. Use of the prediction models would be most appropriate for individuals who resemble the study sample. However, reasonable accuracy in predicting CHD has been demonstrated in the past, when earlier Framingham CHD prediction equations were applied to population samples from Honolulu, Puerto Rico, Albany, Chicago, Los Angeles, Minneapolis, Tecumseh, the Western Collaborative Group, and a national cohort.[59–62] Follow-up from the Framingham Study was also used to estimate CHD experience in men participating in the Multiple Risk Factor Intervention Trial.[63]

Coronary prediction estimates tend to be most reliable when the data are most concentrated and can be particularly useful when subjects have multiple mild abnormalities that act synergistically to increase CHD risk. It is uncommon for persons to have four or five risk factors, and estimates of CHD risk tend to be more precise for individuals with fewer risk factors. Score sheet approaches have been used to target persons for the primary prevention of coronary disease by use of a tabular format called a Sheffield table, in which the estimated absolute risk for CHD is used to establish a threshold for aggressive intervention.[64] The average CHD rates reported in those tables are roughly comparable to the myocardial infarction and coronary death rates among middle-aged men who participated in the West of Scotland trial of cholesterol lowering.[35,65] In contrast, our prediction equations estimate coronary disease risk over a period of 10 years for a larger age range and include total CHD (angina pectoris, myocardial infarction, and coronary death).

A study that considered CHD prediction using TC, LDL-C, TC/HDL-C ratio, and LDL-C/HDL-C ratio[66] concluded that "total cholesterol/HDL is a superior measure of risk for CHD compared with either total cholesterol or LDL cholesterol, and that current practice guidelines could be more efficient if risk stratification was based on this ratio rather than primarily on the LDL cholesterol level." Such an approach appears attractive, but at the extremes of the TC or LDL-C distribution, equal ratios may not signify the same CHD risk. Moreover, use of a ratio may make it harder for the physician to focus on the separate values for TC, LDL-C, and HDL-C that have to be borne in mind to make appropriate clinical decisions concerning therapy. The current approach builds on established blood pressure (JNC-V) and cholesterol (NCEP ATP II) foundations, requires fasting samples only if LDL-C score sheets are used, and is easy to implement as part of a screening program.

Estimation of CHD and other cardiovascular events is a dynamic field. The present formulation has attempted to provide a simplified approach to predict risk for initial CHD events in outpatients free of disease, drawing on national programs for treatment of elevated blood pressure and TC, without a loss in accuracy. Other factors, such as fibrinogen, lipoprotein(a), ERT, family history of premature CHD, and hypertensive therapy have been or will be evaluated as baseline data and greater follow-up experience become available.

TABLE 6. β-Coefficients Underlying CHD Prediction Sheets Using TC Categories

Variable	Men	Women
Age, y	0.04826	0.33766
Age squared, y		−0.00268
TC, mg/dL		
<160	−0.65945	−0.26138
160–199	Referent	Referent
200–239	0.17692	0.20771
240–279	0.50539	0.24385
≥280	0.65713	0.53513
HDL-C, mg/dL		
<35	0.49744	0.84312
35–44	0.24310	0.37796
45–49	Referent	0.19785
50–59	−0.05107	Referent
≥60	−0.48660	−0.42951
Blood pressure		
Optimal	−0.00226	−0.53363
Normal	Referent	Referent
High normal	0.28320	−0.06773
Stage I hypertension	0.52168	0.26288
Stage II–IV hypertension	0.61859	0.46573
Diabetes	0.42839	0.59626
Smoker	0.52337	0.29246
Baseline survival function at 10 years, S(t)	0.90015	0.96246

Appendix

Application of Tables 6 and 7

The β-coefficients given in Table 6 are used to compute a linear function. The latter is corrected for the averages of the participants' risk factors, and the subsequent result is exponentiated and used to calculate a 10-year probability of CHD after insertion into a survival function. The following explanation and an example treat each of these steps in a serial fashion, using Table 6 for the illustration below.

(Equation 1): $\text{L\_Chol}_{\text{men}}=0.04826\times\text{age}-0.65945$ (if cholesterol <160) +0.0 (if cholesterol 160 to 199) +0.17692 (if cholesterol 200 to 239) +0.50539 (if cholesterol 240 to 279) +0.65713 (if cholesterol ≥280) +0.49744 (if HDL-C<35) +0.24310 (if HDL-C 35 to 44) +0.0 (if HDL-C 45 to 49) −0.05107 (if HDL-C 50 to 59) −0.48660 (if HDL-C ≥60) −0.00226 (if blood pressure [BP] optimal) +0.0 (if BP normal) +0.28320 (if BP high normal) +0.52168 (if BP stage I hypertension) +0.61859 (if BP stage II hypertension) +0.42839 (if diabetes present) +0.0 (if diabetes not present) +0.52337 (if smoker) +0.0 (if not smoker).

The function is evaluated at the values of the means for each variable. Call it G, where (Equation 1): $\text{G\_Chol}_{\text{men}}=0.04826\times48.5926-0.65945\times0.07433+0.17692\times0.38851+0.50539\times0.16673+0.65713\times0.05826+$

0.49744×0.19285+0.24310×0.35476−0.05107×0.19646−0.48660×0.10727−0.00226×0.20048+0.28320×0.20048+0.52168×0.22820+0.61859×0.13057+0.42839×0.05223+0.52337×0.40458=3.0975. Similarly, for women, G_Chol=9.92545. For the LDL score sheets, G_LDL for men is 3.00069 and for women 9.914136.

This value of G is subtracted from function L to produce function A (Equation 2), which is then exponentiated, to produce B (Equation 3). The latter represents the relative odds for CHD. The survival value s(t) is exponentiated by B and subtracted from 1.0 to calculate the 10-year probability of CHD (Equation 4).

(Equation 2): A=L−G (where G_Chol=3.0975 for men, 9.92545 for women; similarly for Table 7, G_LDL=3.00069 for men, 9.914136 for women).

(Equation 3): $B=e^{A}$.

(Equation 4): $P=1-[s(t)]^{B}$ [where s(t)_Chol 10 years=0.90015 for men, 0.96246 for women; similarly for Table 7, s(t)_LDL 10 years=0.90017 for men, 0.9628 for women].

Consider a 55-year-old man with cholesterol of 250 mg/dL, HDL-C of 39 mg/dL, blood pressure (146/88 mm Hg) that falls into stage I hypertension, and no diabetes, who is a smoker. In this instance, after Equation 1, L=55×0.04826+0.50539+0.24310+0.52168+0.52337=4.4478. After Equation 2, A=4.4478−3.0975=1.3503, and after Equation 3, $B=e^{1.3503}=3.85874$. Finally, after Equation 4, $P=1-0.90015^{3.85874}=1-0.66637=0.3336$, for a 33% chance of developing CHD over 10 years. According to the point score sheet, 55 years old (4 points)+cholesterol of 250 mg/dL (2 points)+HDL-C of 39 mg/dL (1 point)+stage I blood pressure (2 points)+smoker (2 points)=11 points, corresponding to a 31% chance of developing CHD over 10 years. An average 55-year-old man has a 16% risk, and an ideal man has a 7% risk. Similar calculations can be done for women and for the LDL-C prediction models and score sheets.

TABLE 7. β-Coefficients Underlying CHD Prediction Sheets Using LDL-C Categories

Variable	Men	Women
Age, y	0.04808	0.33994
Age squared, y		−0.0027
LDL-C, mg/dL		
<100	−0.69281	−0.42616
100–129	Referent	Referent
130–159	0.00389	0.01366
160–189	0.26755	0.26948
≥190	0.56705	0.33251
HDL-C, mg/dL		
<35	0.48598	0.88121
35–44	0.21643	0.36312
45–49	Referent	0.19247
50–59	−0.04710	Referent
≥60	−0.34190	−0.35404
Blood pressure		
Optimal	−0.02642	−0.51204
Normal	Referent	Referent
High normal	0.30104	−0.03484
Stage I hypertension	0.55714	0.28533
Stage II–IV hypertension	0.65107	0.50403
Diabetes	0.42146	0.61313
Smoker	0.54377	0.29737
Baseline survival function at 10 years, S(t)	0.90017	0.9628

Acknowledgments

This work is from the National Heart, Lung, and Blood Institute's Framingham Heart Study, supported by NIH/NHLBI contract N01-HC-38038. The authors would like to acknowledge the careful review and helpful criticism by Dr James Cleeman, Coordinator of the National Cholesterol Education Program at the National Heart, Lung, and Blood Institute.

References

1. McGovern PG, Pankow JS, Shahar E, Doliszny KM, Folsom AR, Blackburn H, Luepker RV, the Minnesota Heart Survey Investigators. Recent trends in acute coronary heart disease: mortality, morbidity, medical care, and risk factors. *N Engl J Med.* 1996;334:884–890.
2. Gordon T, Kannel WB. Multiple risk functions for predicting coronary heart disease: the concept, accuracy, and application. *Am Heart J.* 1982; 103:1031–1039.
3. Kannel WB, McGee DL. Diabetes and glucose tolerance as risk factors for cardiovascular disease: the Framingham Study. *Diabetes Care.* 1979;2:120–126.
4. Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR. Diabetes, blood lipids, and the role of obesity in coronary heart disease risk for women. *Ann Intern Med.* 1977;87:393–397.
5. The Expert Panel. Report of the National Cholesterol Education Program Expert Panel on detection, evaluation, and treatment of high blood cholesterol in adults. *Arch Intern Med.* 1988;34:193–201.
6. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Summary of the second report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel II). *JAMA.* 1993;269:3015–3023.
7. The Expert Panel. National Cholesterol Education Program Second Report. The expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel II). *Circulation.* 1994;89:1333–1445.
8. Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J.* 1991;121:293–298.
9. The Expert Panel. Expert panel on detection, evaluation and treatment of high blood cholesterol in adults: summary of the second report of the NCEP expert panel (Adult Treatment Panel II). *JAMA.* 1993;269: 3015–3023.
10. Joint National Committee. The fifth report of the Joint National Committee on detection, evaluation, and treatment of high blood pressure (JNC V). *Arch Intern Med.* 1993;153:154–183.
11. Anderson KM, Wilson PWF, Odell PM, Kannel WB. An updated coronary risk profile: a statement for health professionals. *Circulation.* 1991; 83:357–363.
12. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families: the Framingham Offspring Study. *Am J Epidemiol.* 1979;110:281–290.
13. Abell LL, Levy BB, Brodie BB, Kendall FE. A simplified method for the estimation of total cholesterol in serum and demonstration of its specificity. *J Biol Chem.* 1952;195:357–366.
14. Lipid Research Clinics Program. *Manual of Laboratory Operation.* Bethesda, Md: National Institutes of Health; 1974:75–628.
15. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without the use of the preparative ultracentrifuge. *Clin Chem.* 1972;18:499–502.
16. *Manual of Laboratory Operations: Lipid Research Clinics Program, Lipid and Lipoprotein Analysis.* Washington, DC: National Institutes of Health, US Department of Health and Human Services; 1982.
17. Kannel WB, Wolf PA, Garrison RJ. Monograph *Section 34: Some Risk Factors Related to the Annual Incidence of Cardiovascular Disease and Death Using Pooled Repeated Biennial Measurements: Framingham Heart Study, 30-Year Followup.* Springfield, Mass: National Technical Information Service; 1987:1–459.
18. Neter J, Wasserman W. Multiple regression. In: *Applied Linear Statistical Models.* Homewood, Ill: Irwin; 1974:214–272.
19. Cox DR. Regression models and life tables. *J R Stat Soc B.* 1972;34: 187–220.
20. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 1996;15:361–387.
21. Benfante R, Reed D. Is elevated serum cholesterol level a risk factor for coronary heart disease in the elderly? *JAMA.* 1990;263:393–396.
22. Wilson PWF, Castelli WP, Kannel WB. Coronary risk prediction in adults: the Framingham Heart Study. *Am J Cardiol.* 1987;59:91–94.

23. Corti MC, Guralnik JM, Salive ME, Harris T, Field TS, Wallace RB, Berkman LF, Seeman TE, Glynn RJ, Hennekens CH, Havlik RJ. HDL cholesterol predicts coronary heart disease mortality in older persons. *JAMA.* 1995;274:539–544.
24. Wilson PWF, Kannel WB. Hypercholesterolemia and coronary risk in the elderly: the Framingham Study. *Am J Geriat Cardiol.* 1993;2:52–56.
25. McNamara JR, Cohn JS, Wilson PWF, Schaefer EJ. Calculated values for low-density lipoprotein cholesterol in the assessment of lipid abnormalities and coronary disease risk. *Clin Chem.* 1990;36:36–42.
26. McNamara JR, Cole TG, Contois JH, Ferguson CA, Ordovas JM, Schaefer EJ. Immunoseparation method for measuring low-density lipoprotein cholesterol directly from serum evaluated. *Clin Chem.* 1995; 41:232–240.
27. National Education Programs Working Group report on the management of patients with hypertension and high blood cholesterol. *Ann Intern Med.* 1991;114:224–237.
28. Grover SA, Abrahamowicz M, Joseph L, Brewer C, Coupal L, Suissa S. The benefits of treating hyperlipidemia to prevent coronary heart disease: estimating changes in life expectancy and morbidity. *JAMA.* 1992;267:816–822.
29. Grover SA, Coupal L, Hu XP. Identifying adults at increased risk of coronary disease: how well do the current cholesterol guidelines work? *JAMA.* 1995;274:801–806.
30. Levy D. Have expert panel guidelines kept pace with new concepts in hypertension? *Lancet.* 1995;346:1112.
31. Cooper GR, Myers GL, Smith J, Schlant RC. Blood lipid measurements: variations and practical utility. *JAMA.* 1992;267:1652–1660.
32. Wilson PWF. Cholesterol screening: once is not enough. *Arch Intern Med.* 1995;155:2146–2147.
33. Blankenhorn DH, Nessim SA, Johnson RL, Sanmarco ME, Azen SP, Cashin-Hemphill L. Beneficial effects of combined colestipol-niacin therapy on coronary atherosclerosis and coronary venous bypass grafts. *JAMA.* 1987;257:3233–3240.
34. The 4S Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet.* 1994;344:1383–1389.
35. Shepherd J, Cobbe SM, Ford I, Isles CG, Lorimer AR, MacFarlane PW, McKillop JH, Packard CJ, West of Scotland Coronary Prevention Study Group. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. *N Engl J Med.* 1995;333:1301–1307.
36. Myers RH, Kiely DK, Cupples LA, Kannel WB. Parental history is an independent risk factor for coronary artery disease: the Framingham Study. *Am Heart J.* 1990;120:963–969.
37. Ernst E, Resch KL. Fibrinogen as a cardiovascular risk factor: a meta-analysis and review of the literature. *Ann Intern Med.* 1993;118:956–963.
38. Kannel WB, Wolf R, Castelli WP, D'Agostino RB. Fibrinogen and risk of cardiovascular disease: the Framingham Study. *JAMA.* 1987;258:1183–1186.
39. Kannel WB, D'Agostino RB, Wilson PWF, Belanger AJ, Gagnon DR. Diabetes, fibrinogen, and risk of cardiovascular disease: the Framingham experience. *Am Heart J.* 1990;120:672–676.
40. Barasch E, Benderly M, Graff E, Behar S, Reicher-Reiss H, Caspi A, Pelled B, Reisin L, Roguin N, Goldbourt U. Plasma fibrinogen levels and their correlates in 6457 coronary heart disease patients: the Bezafibrate Infarction Prevention (BIP) Study. *J Clin Epidemiol.* 1995;48:757–765.
41. Pasley BH, Standfast SJ, Katz SH. Prescribing estrogen during menopause: physician survey of practices in 1974 and 1981. *Public Health Rep.* 1984;99:424–429.
42. Bush TL, Cowan LD, Barrett-Connor EL, Criqui MH, Karon JM, Wallace RB, Tyroler HA, Rifkind BM. Estrogen use and all-cause mortality. *JAMA.* 1983;249:903–906.
43. Barrett-Connor EL, Bush TL. Estrogen and coronary heart disease in women. *JAMA.* 1991;265:1861–1867.
44. Stampfer MJ, Colditz GA, Willett WC, Manson JE, Rosner B, Speizer FE, Hennekens CH. Postmenopausal estrogen therapy and cardiovascular disease: ten-year follow-up from the Nurses' Health Study. *N Engl J Med.* 1991;325:756–762.
45. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: quantitative assessment of the epidemiologic evidence. *Prev Med.* 1991;20:47–63.
46. Wilson PWF, Garrison RJ, Castelli WP. Postmenopausal estrogen use, cigarette smoking, and cardiovascular morbidity: the Framingham Study. *N Engl J Med.* 1985;313:1038–1043.
47. Eaker ED, Castelli WP. Coronary heart disease and its risk factors among women in the Framingham Study. In: Eaker ED, Packard B, Wenger NK, Clarkson TB, Tyroler HA, eds. *Coronary Heart Disease in Women.* New York, NY: Haymarket Doyma Inc; 1987:122–130.
48. Petitti DB. Reporting results. In: *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis.* New York, NY: Oxford; 1994:197–211.
49. Powell KE, Thompson PD, Caspersen CJ, Kendrick JS. Physical activity and the incidence of coronary heart disease. *Annu Rev Public Health.* 1987;8:253–287.
50. Lee IM, Hsieh CC, Paffenbarger RS Jr. Exercise intensity and longevity in men: the Harvard Alumni Health Study. *JAMA.* 1995;273:1179–1184.
51. Berlin JA, Colditz GA. A meta-analysis of physical activity in the prevention of coronary heart disease. *Am J Epidemiol.* 1990;132:612–628.
52. Wilson PWF. High-density lipoprotein, low-density lipoprotein and coronary artery disease. *Am J Cardiol.* 1990;66(suppl A):7–10.
53. Anderson KM, Wilson PWF, Garrison RJ, Castelli WP. Longitudinal and secular trends in lipoprotein cholesterol measurements in a general population sample: the Framingham Offspring Study. *Atherosclerosis.* 1987;68:59–66.
54. Helmrich SP, Ragland DR, Leung RW, Paffenbarger RS Jr. Physical activity and reduced occurrence of non-insulin-dependent diabetes mellitus. *N Engl J Med.* 1991;325:147–152.
55. Burchfiel CM, Curb JD, Sharp DS, Rodriguez BL, Arakaki R, Chyou PH, Yano K. Distribution and correlates of insulin in elderly men: the Honolulu Heart Program. *Arterioscler Thromb Vasc Biol.* 1995;15:2213–2221.
56. Wood PD. Physical activity, diet, and health: independent and interactive effects. *Med Sci Sports Exerc.* 1994;26:838–843.
57. Dannenberg AL, Keller JB, Wilson PWF, Castelli WP. Leisure time physical activity in the Framingham Offspring Study: description, seasonal variation, and risk factor correlates. *Am J Epidemiol.* 1989;129:76–87.
58. Wood PD, Haskell WL, Klein H, Lewis S, Stern MP, Farquhar JW. The distribution of plasma lipoproteins in middle-aged male runners. *Metabolism.* 1976;25:1249–1257.
59. Gordon T, Garcia-Palmieri MR, Kagan A, Kannel WB, Schiffman J. Differences in coronary heart disease in Framingham, Honolulu and Puerto Rico. *J Chronic Dis.* 1974;27:329–344.
60. McGee D, T Gordon. *The Framingham Study applied to four other U. S. based epidemiological studies of cardiovascular disease (Section No. 31).* Bethesda, Md: US Department of Health, Education, and Welfare, NIH; 1976:76–1083.
61. Brand RJ, Rosenman RH, Scholtz RI. Multivariate prediction of coronary heart disease in the Western Collaborative Group Study compared to the findings of the Framingham Study. *Circulation.* 1976;53:348–355.
62. Leaverton PE, Sorlie PD, Kleinman JC, Dannenberg AL, Ingster-Moore L, Kannel WB, Cornoni-Huntley JC. Representativeness of the Framingham risk model for coronary heart disease mortality: a comparison with a national cohort study. *J Chronic Dis.* 1987;40:775–784.
63. The Multiple Risk Factor Intervention Trial Group. Statistical design considerations in the NHLI multiple risk factor intervention trial (MRFIT). *J Chronic Dis.* 1977;30:261–275.
64. Ramsay LE, Haq IU, Jackson PR, Yeo WW, Pickin DM, Payne JN. Targeting lipid-lowering drug therapy for primary prevention of coronary disease: an updated Sheffield table. *Lancet.* 1996;348:387–388.
65. West of Scotland Coronary Prevention Group. West of Scotland Coronary Prevention Study: identification of high-risk groups and comparison with other cardiovascular intervention trials. *Lancet.* 1996;348:1339–1342.
66. Kinosian B, Glick H, Garland G. Cholesterol and coronary heart disease: predicting risks by levels and ratios. *Ann Intern Med.* 1994;121:641–647.

Genome-wide association study identifies novel breast cancer susceptibility loci

Douglas F. Easton[1], Karen A. Pooley[2], Alison M. Dunning[2], Paul D. P. Pharoah[2], Deborah Thompson[1], Dennis G. Ballinger[3], Jeffery P. Struewing[4], Jonathan Morrison[2], Helen Field[2], Robert Luben[5], Nicholas Wareham[5], Shahana Ahmed[2], Catherine S. Healey[2], Richard Bowman[6], the SEARCH collaborators[2*], Kerstin B. Meyer[7], Christopher A. Haiman[8], Laurence K. Kolonel[9], Brian E. Henderson[8], Loic Le Marchand[9], Paul Brennan[10], Suleeporn Sangrajrang[11], Valerie Gaborieau[10], Fabrice Odefrey[10], Chen-Yang Shen[12], Pei-Ei Wu[12], Hui-Chun Wang[12], Diana Eccles[13], D. Gareth Evans[14], Julian Peto[15], Olivia Fletcher[16], Nichola Johnson[16], Sheila Seal[17], Michael R. Stratton[17,18], Nazneen Rahman[17], Georgia Chenevix-Trench[19], Stig E. Bojesen[20], Børge G. Nordestgaard[20], Christen K. Axelsson[21], Montserrat Garcia-Closas[22], Louise Brinton[22], Stephen Chanock[23], Jolanta Lissowska[24], Beata Peplonska[25], Heli Nevanlinna[26], Rainer Fagerholm[26], Hannaleena Eerola[26,27], Daehee Kang[28], Keun-Young Yoo[28,29], Dong-Young Noh[28], Sei-Hyun Ahn[30], David J. Hunter[31,32], Susan E. Hankinson[32], David G. Cox[31], Per Hall[33], Sara Wedren[33], Jianjun Liu[34], Yen-Ling Low[34], Natalia Bogdanova[35,36], Peter Schürmann[36], Thilo Dörk[36], Rob A. E. M. Tollenaar[37], Catharina E. Jacobi[38], Peter Devilee[39], Jan G. M. Klijn[40], Alice J. Sigurdson[41], Michele M. Doody[41], Bruce H. Alexander[42], Jinghui Zhang[4], Angela Cox[43], Ian W. Brock[43], Gordon MacPherson[43], Malcolm W. R. Reed[44], Fergus J. Couch[45], Ellen L. Goode[45], Janet E. Olson[45], Hanne Meijers-Heijboer[46,47], Ans van den Ouweland[47], André Uitterlinden[48], Fernando Rivadeneira[48], Roger L. Milne[49], Gloria Ribas[49], Anna Gonzalez-Neira[49], Javier Benitez[49], John L. Hopper[50], Margaret McCredie[51], Melissa Southey[50], Graham G. Giles[52], Chris Schroen[53], Christina Justenhoven[54], Hiltrud Brauch[54], Ute Hamann[55], Yon-Dschun Ko[56], Amanda B. Spurdle[19], Jonathan Beesley[19], Xiaoqing Chen[19], kConFab[57*], AOCS Management Group[19,57*], Arto Mannermaa[58,59], Veli-Matti Kosma[58,59], Vesa Kataja[58,60], Jaana Hartikainen[58,59], Nicholas E. Day[5], David R. Cox[3] & Bruce A. J. Ponder[2,7]

Breast cancer exhibits familial aggregation, consistent with variation in genetic susceptibility to the disease. Known susceptibility genes account for less than 25% of the familial risk of breast cancer, and the residual genetic variance is likely to be due to variants conferring more moderate risks. To identify further susceptibility alleles, we conducted a two-stage genome-wide association study in 4,398 breast cancer cases and 4,316 controls, followed by a third stage in which 30 single nucleotide polymorphisms (SNPs) were tested for confirmation in 21,860 cases and 22,578 controls from 22 studies. We used 227,876 SNPs that were estimated to correlate with 77% of known common SNPs in Europeans at $r^2 > 0.5$. SNPs in five novel independent loci exhibited strong and consistent evidence of association with breast cancer ($P < 10^{-7}$). Four of these contain plausible causative genes (*FGFR2, TNRC9, MAP3K1* and *LSP1*). At the second stage, 1,792 SNPs were significant at the $P < 0.05$ level compared with an estimated 1,343 that would be expected by chance, indicating that many additional common susceptibility alleles may be identifiable by this approach.

Breast cancer is about twice as common in the first-degree relatives of women with the disease as in the general population, consistent with variation in genetic susceptibility to the disease[1]. In the 1990s, two major susceptibility genes for breast cancer, *BRCA1* and *BRCA2*, were identified[2,3]. Inherited mutations in these genes lead to a high risk of breast and other cancers[4]. However, the majority of multiple case breast cancer families do not segregate mutations in these genes. Subsequent genetic linkage studies have failed to identify further major breast cancer genes[5]. These observations have led to the proposal that breast cancer susceptibility is largely 'polygenic': that is, susceptibility is conferred by a large number of loci, each with a small effect on breast cancer risk[6]. This model is consistent with the observed patterns of familial aggregation of breast cancer[7]. However, progress in identifying the relevant loci has been slow. As linkage studies lack power to detect alleles with moderate effects on risk, large case-control association studies are required. Such studies have identified variants in the DNA repair genes *CHEK2, ATM, BRIP1* and *PALB2* that confer an approximately twofold risk of breast cancer, but these variants are rare in the population[8–14]. A recent study has shown that a common coding variant in *CASP8* is associated with a moderate reduction in breast cancer risk[15]. After accounting for all the known breast cancer loci, more than 75% of the familial risk of the disease remains unexplained[16].

Recent technological advances have provided platforms that allow hundreds of thousands of SNPs to be analysed in association studies, thus providing a basis for identifying moderate risk alleles without

Affiliations of the above authors are given at the end of the paper.
*Lists of consortia participants and affiliations appear after author affiliations.

prior knowledge of position or function. It has been estimated that there are 7 million common SNPs in the human genome (with minor allele frequency, m.a.f., >5%)[17]. However, because recombination tends to occur at distinct 'hot-spots', neighbouring polymorphisms are often strongly correlated (in 'linkage disequilibrium', LD) with each other. The majority of common genetic variants can therefore be evaluated for association using a few hundred thousand SNPs as tags for all the other variants[18]. We aimed to identify further breast cancer susceptibility loci in a three-stage association study[19]. In the first stage, we used a panel of 266,722 SNPs, selected to tag known common variants across the entire genome[18]. These SNPs were genotyped in 408 breast cancer cases and 400 controls from the UK; data were analysed for 390 cases and 364 controls genotyped for ≥80% of the SNPs. The cases were selected to have a strong family history of breast cancer, equivalent to at least two affected female first-degree relatives, because such cases are more likely to carry susceptibility alleles[20]. Initally, we analysed 227,876 SNPs (85%) with genotypes on at least 80% of the subjects. We estimate that these SNPs are correlated with 58% of common SNPs in the HapMap CEPH/CEU (Utah residents with ancestry from northern and western Europe) samples at $r^2 > 0.8$, and 77% at $r^2 > 0.5$ (mean $r^2 = 0.75$; see Supplementary Fig. 1) (http://www.hapmap.org/)[21]. As expected, coverage was strongly related to m.a.f.: 70% of SNPs with m.a.f. > 10% were tagged at $r^2 > 0.8$, compared with 23% of SNPs with m.a.f. 5–10%. The main analyses were restricted to 205,586 SNPs that had a call rate of 90% and whose genotype distributions did not differ from Hardy–Weinberg equilibrium in controls (at $P < 10^{-5}$).

For the second stage we selected 12,711 SNPs, approximately 5% of those typed in stage 1, on the basis of the significance of the difference in genotype frequency between cases and controls. These SNPs were then genotyped in a further 3,990 invasive breast cancer cases and 3,916 controls from the SEARCH study, using a custom-designed oligonucleotide array. In the main analyses, we considered 10,405 SNPs with call rate of >95% that did not deviate from Hardy–Weinberg equilibrium in controls.

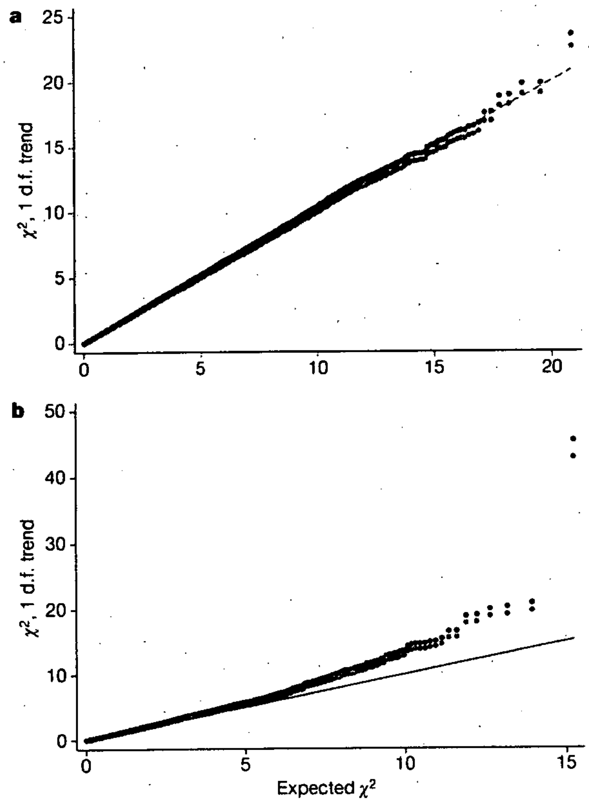


Figure 1 | Quantile-quantile plots for the test statistics (Cochran-Armitage 1 d.f. χ^2 trend tests) for stages 1 and 2. a, Stage 1; **b**, stage 2. Black dots are the uncorrected test statistics. Red dots are the statistics corrected by the genomic control method ($\lambda = 1.03$ for stage 1, $\lambda = 1.06$ for stage 2). Under the null hypothesis of no association at any locus, the points would be expected to follow the black line.

Comparison of the observed and expected distribution of test statistics showed some evidence for an inflation of the test statistics in both stage 1 (inflation factor $\lambda = 1.03$, 95% confidence interval (CI) 1.02–1.04) and stage 2 ($\lambda = 1.06$, 95% CI 1.04–1.12), based on the 90% least significant SNPs (Fig. 1). Possible explanations for this inflation include population stratification, cryptic relatedness among subjects, and differential genotype calling between cases and controls. There was evidence for an excess of low call rate SNPs among the most significant SNPs ($P < 0.01$) in stage 1, but not in stage 2, suggesting that some of this effect is a genotyping artefact (Supplementary Table 1). However, the inflation was still present among SNPs with call rate >99% in both cases and controls, possibly reflecting population substructure. We computed 1 degree of freedom (d.f.) association tests for each SNP, combining stages 1 and 2. After adjustment for this inflation by the genomic control method[22], we observed more associations than would have been expected by chance at $P < 0.05$ (Table 1). One SNP (dbSNP rs2981582) was significant at the $P < 10^{-7}$ level that has been proposed as appropriate for genome-wide studies[23].

In the third stage, to establish whether any SNPs were definitely associated with risk, we tested 30 of the most significant SNPs in 22 additional case-control studies, comprising 21,860 cases of invasive breast cancer, 988 cases of carcinoma *in situ* (CIS) and 22,578 controls (Supplementary Table 2). Six SNPs showed associations in stage 3 that were significant at $P \leq 10^{-5}$ with effects in the same direction as in stages 1 and 2 (Table 2, Supplementary Table 3, and Fig. 2). All these SNPs reached a combined significance level of $P < 10^{-7}$ (ranging from 2×10^{-76} to 3×10^{-9}). Of these six SNPs, five were within genes or LD blocks containing genes. SNP rs2981582 lies in intron 2 of *FGFR2* (also known as *CEK3*), which encodes the fibroblast growth factor receptor 2. SNPs rs12443621 and rs8051542 are both located in an LD block containing the 5′ end of *TNRC9* (also known as *TOX3*), a gene of uncertain function containing a tri-nucleotide repeat motif, as well as the hypothetical gene, *LOC643714*. SNP rs889312 lies in an LD block of approximately 280 kb that contains *MAP3K1* (also known as *MEKK*), which encodes the signalling protein mitogen-activated protein kinase kinase kinase 1, in addition to two other genes: *MGC33648* and *MIER3*. SNP rs3817198 lies in intron 10 of *LSP1* (also known as *WP43*), encoding lymphocyte-specific protein 1, an F-actin bundling cytoskeletal protein expressed in haematopoietic and endothelial cells. A further SNP, rs2107425, located just 110 kilobases (kb) from rs3817198, was also identified (overall $P = 0.00002$). rs2107425 is within the *H19* gene, an imprinted maternally expressed untranslated messenger RNA closely involved in regulation of the insulin growth factor gene, *IGF2*. In stage 3, however, rs2107425 was only weakly significant after adjustment for rs3817198 by logistic regression ($P = 0.06$). This suggests that the association with breast cancer risk may be driven by variants in *LSP1* rather than in *H19*. The sixth SNP reaching a combined $P < 10^{-7}$ was rs13281615, which lies on 8q. It is correlated with SNPs in a 110 kb LD block that contains no known

Table 1 | Number of significant associations after stage 2

Level of significance	Observed	Observed adjusted*	Expected	Ratio
0.01–0.05	1,239	1,162	934.3	1.24
0.001–0.01	574	517	347.6	1.49
0.0001–0.001	112	88	53.3	1.65
0.00001–0.0001	16	12	7.0	1.71
<0.00001	15	13	0.96	13.5
All $P < 0.05$	1,956	1,792	1,343.2	1.33

Observed numbers of SNPs associated with breast cancer after stage 2, by level of significance, before and after adjustment for population stratification, and expected numbers under the null hypothesis of no association.

* Adjusted for inflation of the test statistic by the genomic control method.

Table 2 | Summary of results for eleven SNPs selected for stage 3 that showed evidence of an association with breast cancer

rs Number	Gene	Position*	m.a.f.†	Per allele OR (95% CI)	HetOR (95% CI)	HomOR (95% CI)	*P*-trend		
							Stages 1 and 2	Stage3	Combined
rs2981582	*FGFR2*	10q 123342307	0.38 *(0.30)*	1.26 (1.23–1.30)	1.23 (1.18–1.28)	1.63 (1.53–1.72)	4×10^{-16}	5×10^{-62}	2×10^{-76}
rs12443621	*TNRC9/* *LOC643714*	16q 51105538	0.46 *(0.60)*	1.11 (1.08–1.14)	1.14 (1.09–1.20)	1.23 (1.17–1.30)	10^{-7}	9×10^{-14}	2×10^{-19}
rs8051542	*TNRC9/* *LOC643714*	16q 51091668	0.44 *(0.20)*	1.09 (1.06–1.13)	1.10 (1.05–1.16)	1.19 (1.12–1.27)	4×10^{-6}	4×10^{-8}	10^{-12}
rs889312	*MAP3K1*	5q 56067641	0.28 *(0.54)*	1.13 (1.10–1.16)	1.13 (1.09–1.18)	1.27 (1.19–1.36)	4×10^{-6}	3×10^{-15}	7×10^{-20}
rs3817198	*LSP1*	11p 1865582	0.30 *(0.14)*	1.07 (1.04–1.11)	1.06 (1.02–1.11)	1.17 (1.08–1.25)	8×10^{-6}	10^{-5}	3×10^{-9}
rs2107425	*H19*	11p 1977651	0.31 *(0.44)*	0.96 (0.93–0.99)	0.94 (0.90–0.98)	0.95 (0.89–1.01)	7×10^{-6}	0.01	2×10^{-5}
rs13281615		8q 128424800	0.40 *(0.56)*	1.08 (1.05–1.11)	1.06 (1.01–1.11)	1.18 (1.10–1.25)	2×10^{-7}	6×10^{-7}	5×10^{-12}
rs981782		5p 45321475	0.47 *(0.37)*	0.96 (0.93–0.99)	0.96 (0.92–1.01)	0.92 (0.87–0.97)	8×10^{-5}	0.003	9×10^{-6}
rs30099		5q 52454339	0.08 *(0.39)*	1.05 (1.01–1.10)	1.06 (1.00–1.11)	1.09 (0.96–1.24)	0.003	0.02	0.001
rs4666451		2p 19150424	0.41 *(0.04)*	0.97 (0.94–1.00)	0.98 (0.93–1.02)	0.93 (0.87–0.99)	5×10^{-6}	0.04	6×10^{-5}
rs3803662‡	*TNRC9/* *LOC643714*	16q 51143842	0.25 *(0.60)*	1.20 (1.16–1.24)	1.23 (1.18–1.29)	1.39 (1.26–1.45)	3×10^{-12}	10^{-26}	10^{-36}

OR, odds ratio; HetOR, odds ratio in heterozygotes; HomOR, odds ratio in rare homozygotes (relative to common homozygotes); CI, confidence interval.

* Build 36.2 position.

† Minor allele frequency in SEARCH (UK) study. Combined allele frequency from three Asian studies in italics.

‡ rs3803662 was not part of the initial tag SNP set but identified as a result of fine-scale mapping of the *TNRC9/LOC643714* locus and typed in the stage 2 and stage 3 sets (but not the stage 1 set).

genes. The basis of this association therefore remains obscure. This SNP is approximately 130 kb proximal to rs1447295, 60 kb proximal to rs6983267 and 230 kb distal to rs16901979, recently shown to be associated with prostate cancer[24–26].

In addition to the seven SNPs described above, there was evidence of association among the remaining 23 SNPs (global $P = 0.001$ in stage 3). In particular, three SNPs showed some evidence of association in stage 3 ($P < 0.05$, in each case in the same direction as in stages 1 and 2; Table 2). SNPs rs981782 and rs30099 both lie in the centromeric region of chromosome 5. rs4666451 lies on 2p, a region for which some evidence of linkage to breast cancer in families has been reported[5]. The 20 other SNPs showed no evidence of association in stage 3 (global $P = 0.11$), suggesting that most of these associations from stages 1 and 2 were false positives.

FGFR2

The most significantly associated SNP, rs2981582, lies within a 25 kb LD block almost entirely within intron 2 of *FGFR2*. We found no evidence of association with SNPs elsewhere in the gene (Fig. 3a). In an attempt to identify a causal variant, we first identified the 19 common variants (m.a.f. > 0.05) in this block from HapMap CEU data. These were tagged ($r^2 > 0.8$) by 7 SNPs including rs2981582. The additional tag SNPs were genotyped in the SEARCH study cases and controls. Multiple logistic regression analysis of these variants found no additional evidence for association after adjusting for rs2981582. Haplotype analysis of these 7 SNPs indicated that multiple haplotypes carrying the minor (*a*) allele of rs2981582 were associated with an increased risk of breast cancer, implying that the association was being driven by rs2981582 itself or a variant strongly correlated with it (Supplementary Table 4).

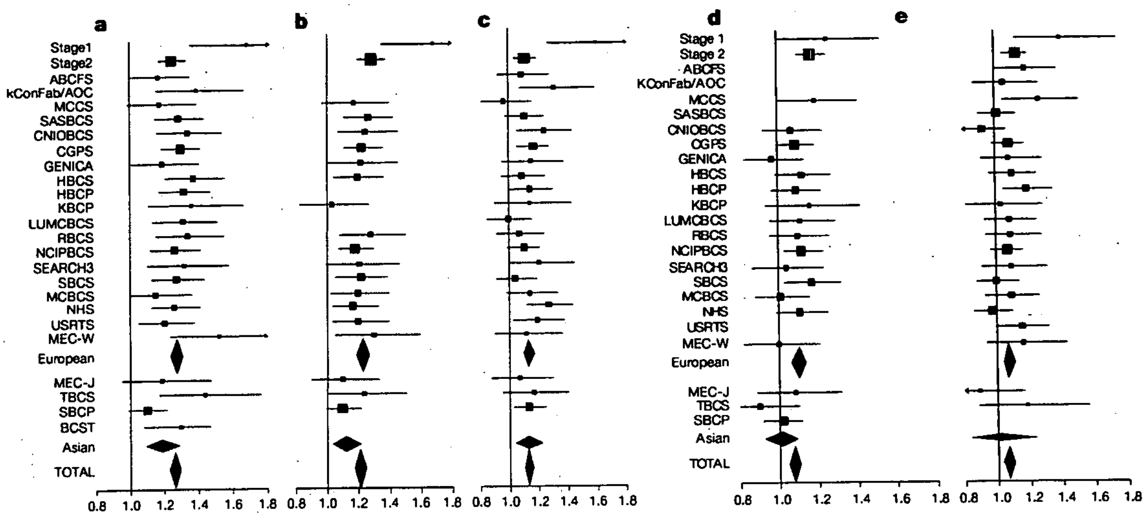


Figure 2 | Forest plots of the per-allele odds ratios for each of the five SNPs reaching genome-wide significance. a, rs2981582; **b**, rs3803662; **c**, rs889312; **d**, rs13281615; and **e**, rs3817198. The *x*-axis gives the per-allele odds ratio. Each row represents one study (see Supplementary Table 2), with summary odds ratios for all European and all Asian studies, and all studies combined. The area of the square for each study is proportional to the inverse of the variance of the estimate. Horizontal lines represent 95% confidence intervals. Diamonds represent the summary odds ratios, with 95% confidence intervals, based on the stage 3 studies only.

Resequencing of this region in 45 subjects of European origin identified 29 variants that were strongly correlated with rs2981582 ($r^2 > 0.6$) (http://cgwb.nci.nih.gov; Fig. 3b and Supplementary Tables 5–8). A subset of 14 variants tagged 27 of these in European ($r^2 > 0.95$) and Asian (Korean) samples ($r^2 > 0.86$). Two variants could not be genotyped reliably. This new tagging set was then genotyped in SEARCH and 3 studies from Asian populations; the Asian studies were included because the LD is weaker, providing greater power to resolve the causal variant (Fig. 3b, left panel). The strongest association was found with rs7895676. On the assumption that there is a single disease-causing allele, we calculated a likelihood for each variant. 21 SNPs (including rs2981582) had a likelihood ratio of <1/100 relative to rs7895676, indicating that none of these are likely to be the causal variant (Supplementary Table 8). Six variants were too strongly correlated for their individual effects to be separated using a genetic epidemiological approach. Functional assays will be required to determine which is causally related to breast cancer risk.

Intron 2 of *FGFR2* shows a high degree of conservation in mammals, and contains several putative transcription-factor binding sites (http://genomequebec.mcgill.ca/PReMod)[27], some of which lie in close proximity to the relevant SNPs. We therefore speculate that the association with breast cancer risk is mediated through regulation of *FGFR2* expression. Of possible relevance is that only three of these variants (rs10736303, rs2981578 and rs35054928) are within sequences conserved across all placental mammals (Fig. 3c and Supplementary Table 8). Of these, the disease associated allele of rs10736303 generates a putative oestrogen receptor (ER) binding site. rs35054928 lies immediately adjacent to a perfect POU domain protein octamer (Oct) binding site. However, multiple splice variants have been reported in *FGFR2*, and differential splicing might provide an alternative mechanism for the association. FGFR2 is a receptor tyrosine kinase that is amplified and overexpressed in 5–10% of breast tumours[28–30]. Somatic missense mutations of *FGFR2* that are likely to be implicated in cancer development have also been demonstrated in primary tumours and cell lines of multiple tumour types (http://www.sanger.ac.uk/genetics/CGP/cosmic/)[30,31].

TNRC9/LOC643714 locus

As two SNPs in the *TNRC9/LOC643714* locus, rs12443621 and rs8051542, both showed convincing evidence of association, we further evaluated this region by genotyping, in the SEARCH set, an additional 19 SNPs tagging 101 common variants within the entire *TNRC9* and *LOC643714* genes, based on the HapMap CEU data. SNPs tagging the coding region of *TNRC9* showed no evidence of association. The strongest association was observed with rs3803662, a synonymous coding SNP of *LOC643714* that lies 8 kb upstream of *TNRC9*. This SNP was therefore genotyped in the stage 3 set (Table 2). Logistic regression analysis indicated that rs3803662 exhibited a stronger association with disease than other SNPs, and the associations with other SNPs were non-significant after adjustment for rs3803662. These results suggest

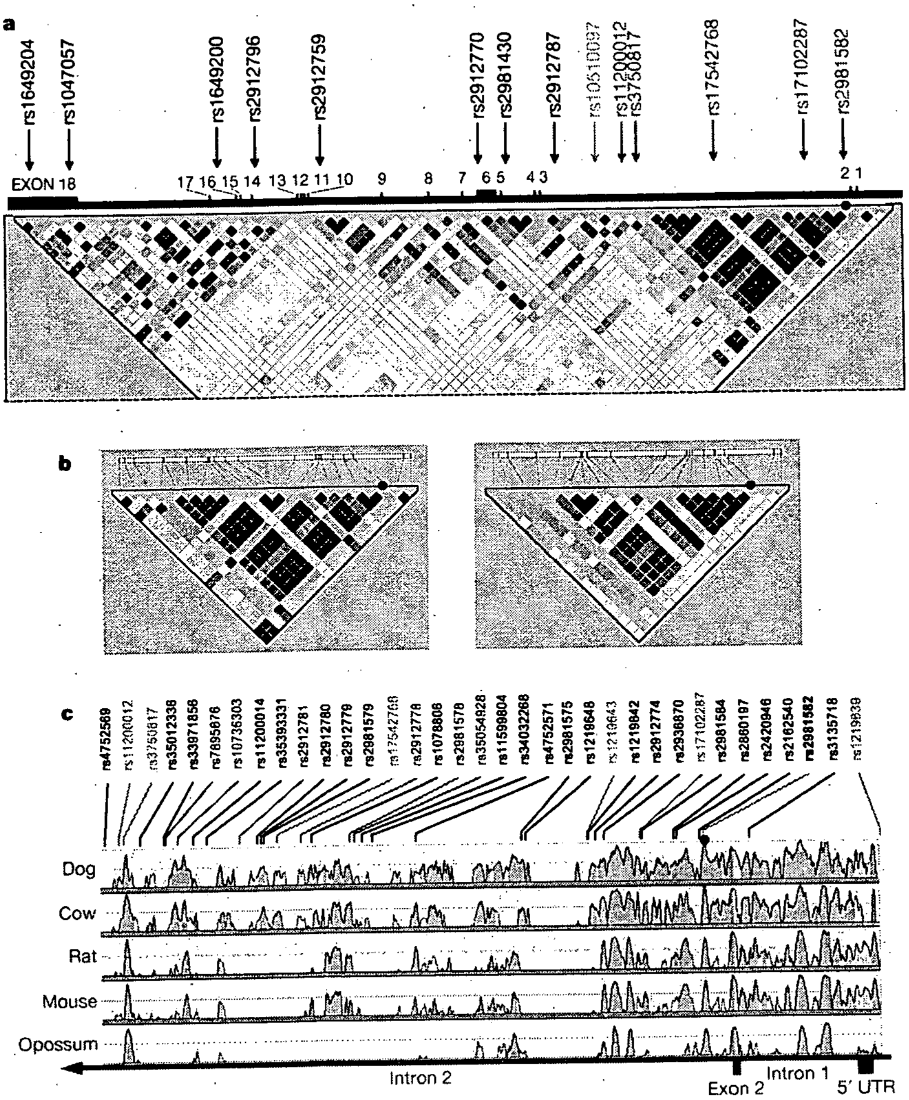


Figure 3 | The *FGFR2* locus. **a**, Map of the whole *FGFR2* gene, viewed relative to common SNPs on HapMap. The gene is 126 kb long and in reverse 3′–5′ orientation on chromosome 10. Exon positions are illustrated with respect to the 67 SNPs with m.a.f. > 5% in HapMap CEU (therefore the map is not to physical scale). Numbered SNPs are those tested in the genome-wide study. SNPs in black were not significant in stage 1. Those in red were significant at $P < 0.0001$ after stage 2. rs10510097 (orange) was significant in stage 1, but failed quality control in stage 2 owing to deviation from Hardy–Weinberg equilibrium. Squares indicate pairwise r^2 on a greyscale (black = 1, white = 0). Red circle indicates rs2981582. **b**, Resequenced 32 kb region, shown relative to SNPs in CEU with m.a.f. > 5%, showing pairwise LD for SNPs in HapMap CEU (left panel) and JPT/CHB (right panel). Red circle indicates rs2981582, shown in bold black. **c**, Sequence conservation of 32 kb region in five species, relative to human sequence (http://pipeline.lbl.gov/methods.shtml)[35]. Red circle indicates rs2981582. SNPs in grey are those used in the initial tagging of known common HapMap SNPs within the block. SNPs in black are correlated with rs2981582 with $r^2 > 0.6$ in European samples. Six SNPs in red were those consistent with being the causative variant on the basis of the genetic data (not excluded at odds of 100:1 relative to the SNP with the strongest association, rs7895676).

that the causal variant is closely correlated with rs3803662. Four SNPs in the HapMap CEU data (rs17271951, rs11362548, rs3095604 and rs4784227) that span *LOC643714* and the 5′ regulatory regions of *TNRC9* are strongly correlated with rs3803662, and it therefore remains unclear in which gene the causative variant lies. *TNRC9* contains a putative HMG (high mobility group) box motif, suggesting that it might act as a transcription factor.

Pattern of risks

We assessed in more detail, in the stage 3 data, the pattern of the risks associated with the five independent SNPs that reached an overall $P < 10^{-7}$: rs2981582 (*FGFR2*), rs3803662 (*TNRC9/LOC643714*), rs889312 (*MAP3K1*), rs13281615 (8q) and rs3817198 (*LSP1*). For each of these five SNPs, the minor allele in Europeans was associated with an increased risk of breast cancer in a dose-dependent manner, with a higher risk of breast cancer in homozygous than in heterozygous carriers. Simple dominant and recessive models could be rejected for each SNP (all $P = 0.02$ or less). There was a marked difference in allele frequencies between populations, with the risk-associated alleles of rs8051542, rs889312 and rs13281615 being the major allele in Asian populations. The per allele odds ratio associated with rs2981582 was significantly smaller, though still elevated, in the Asian versus European populations ($P = 0.04$ for difference in odds ratio). This difference is consistent with the hypothesis that rs2981582 is not the functional variant at the *FGFR2* locus, and was not seen for SNPs exhibiting stronger evidence in the fine-scale mapping. No other evidence for heterogeneity in the per-allele odds ratio among studies was observed (Fig. 2).

Three of the SNPs (rs2981582, rs3803662 and rs889312) also showed evidence of association with breast CIS (Supplementary Table 9). For rs2981582 and rs3803662, the estimated odds ratios were greater for a diagnosis of breast cancer before age 40 years, but the trends by age were not statistically significant (Supplementary Table 10). There was evidence of an association with family history of breast cancer for three SNPs: for rs2981582 ($P = 0.02$), rs3803662 ($P = 0.03$) and rs13281615 ($P = 0.05$), the susceptibility allele was commoner in women with a first-degree relative with the disease than in those without (Supplementary Table 11). rs2981582 was also associated with bilaterality ($P = 0.02$). The associations with family history and bilaterality are to be expected for susceptibility loci, and are similar to previous observations for alleles in *CHEK2* and *ATM* (refs 10, 12, 14).

Discussion

This study has identified five novel breast cancer susceptibility loci, and demonstrated conclusively that some of the variation in breast cancer risk is due to common alleles. None of the loci we identified had been previously reported in association studies. Most previously identified breast cancer susceptibility genes are involved in DNA repair, and many association studies in breast cancer have concentrated on genes in DNA repair and sex hormone synthesis and metabolism pathways. None of the associations reported here appear to relate to genes in these pathways. It is notable that three of the five loci contain genes related to control of cell growth or to cell signalling, but only one (*FGFR2*) had a clear prior relevance to breast cancer. These results should, therefore, open up new avenues for basic research.

Our results emphasize the critical importance of study size in genetic association studies. It is notable that none of the confirmed associations reached genome-wide significance after stage 1 and only one reached this level after stage 2. As most common cancers have similar familial relative risks to breast cancer, it is likely that similarly large studies will be required to identify common alleles for other cancers. The fine-scale mapping of the *FGFR2* locus demonstrates that, even with a clear association, identification of the causative variant can be extremely problematic. However, the use of studies from multiple populations with different patterns of LD can substantially reduce the number of variants that need to be subjected to functional analysis.

As these susceptibility alleles are very common, a high proportion of the general population are carriers of at-risk genotypes. For example, approximately 14% of the UK population and 19% of UK breast cancer cases are homozygous for the rare allele at rs2981582. On the other hand, the increased risks associated with these alleles are relatively small—on the basis of UK population rates, the estimated breast cancer risk by age 70 years for rare homozygotes at rs2981582 is 10.5%, compared to 6.7% in heterozygotes and 5.5% in common homozygotes. At this stage, it is unlikely that these SNPs will be appropriate for predictive genetic testing, either alone or in combination with each other. However, as further susceptibility alleles are identified, a combination of such alleles together with other breast cancer risk factors may become sufficiently predictive to be important clinically.

On the basis of the relative risk estimates from stage 3, and assuming that the five most significant loci interact multiplicatively on disease risk, these loci explain an estimated 3.6% of the excess familial risk of breast cancer. On the basis of our staged design and the estimated distribution of linkage disequilibrium between the typed SNPs and those in HapMap, we estimate that the power to identify the five most significant associations at $P < 10^{-7}$ (rs2981582, rs3803662, rs889312, rs13281615 and rs3817198) was 93%, 71%, 25%, 3% and 1% respectively. These estimates are uncertain, notably because the true coverage of HapMap SNPs is unknown. Nevertheless, these calculations indicate that the power to detect the two strongest associations was high, and suggest that there are likely to be few other common variants with a similar effect on variation in breast cancer risk to rs2981582. In contrast, the low power to detect rs13281615 and rs3817198 suggests that these variants may represent a much larger class of loci, each explaining of the order of 0.1% of the familial risk of breast cancer. An example of such a locus is provided by *CASP8* D302H, which showed strong evidence of association in a previous large study[15]. This SNP was tested in stage 1, but the association was missed because it did not reach the threshold for testing in stage 2. The excess of associations after stage 2 is also consistent with the existence of many such loci. In addition, because the coverage for SNPs with m.a.f. < 10% was low, many low frequency alleles may have been missed. The detection of further susceptibility loci will require genome-wide studies with more complete coverage and using larger numbers of cases and controls, together with the combination of results across multiple studies. The present study demonstrates that common susceptibility loci can be reliably identified, and that they may together explain an appreciable fraction of the genetic variance in breast cancer risk.

METHODS SUMMARY

Cases for stage 1 were identified through clinical genetics centres in the UK and a national study of bilateral breast cancer. Cases in stage 2 were drawn from a population-based study of breast cancer (SEARCH)[32]. Controls for stages 2 and 3 were drawn from EPIC-Norfolk, a population-based study of diet and cancer[33].

Cases and controls for stage 3 were identified through case-control studies in Europe, North America, South-East Asia and Australia participating in the Breast Cancer Association Consortium (Supplementary Table 2)[34].

Genotyping for stages 1 and 2 was conducted using high-density oligonucleotide microarrays. For the main analyses, we excluded samples called on ≤80% of SNPs in either stage. We also excluded SNPs that achieved a call rate of ≤90% in stage 1 and ≤95% in stage 2, and SNPs whose frequency deviated from Hardy–Weinberg equilibrium in controls at $P < 0.00001$. Genotyping for stage 3, and for the fine-scale mapping of the *FGFR2* locus, was conducted using either a 5′ nuclease assay (Taqman, Applied Biosystems) or MALDI-TOF mass spectrometry using the Sequenom iPLEX system. For each centre, we excluded any sample called on ≤80% of SNPs, and any SNP with a call rate of ≤95% or a deviation from Hardy–Weinberg equilibrium in controls at $P < 0.00001$. Tests of association were 1 d.f. Cochran-Armitage tests, stratified for stage, centre and ethnic group (European or Asian). Odds ratios for each SNP were estimated using stratified logistic regression, using the stage 3 data only.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 February; accepted 30 April 2007.
Published online 27 May 2007; corrected 28 June 2007 (details online).

1. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: Collaborative reanalysis of individual data from 52 epidemiological

studies including 58 209 women with breast cancer and 101 986 women without the disease. *Lancet* 358, 1389–1399 (2001).

2. Miki, Y. *et al.* A strong candidate for the breast and ovarian-cancer susceptibility gene BRCA1. *Science* 266, 66–71 (1994).
3. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* 378, 789–792 (1995).
4. Antoniou, A. *et al.* Average risks of breast and ovarian cancer associated with mutations in *BRCA1* or *BRCA2* detected in case series unselected for family history: A combined analysis of 22 studies. *Am. J. Hum. Genet.* 72, 1117–1130 (2003).
5. Smith, P. *et al.* A genome wide linkage search for breast cancer susceptibility genes. *Genes Chromosom. Cancer* 45, 646–655 (2006).
6. Pharoah, P. D. P. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genet.* 31, 33–36 (2002).
7. Antoniou, A. C., Pharoah, P. D. P., Smith, P. & Easton, D. F. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *Br. J. Cancer* 91, 1580–1590 (2004).
8. Rahman, N. *et al. PALB2*, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nature Genet.* 39, 165–167 (2007).
9. Thompson, D. *et al.* Cancer risks and mortality in heterozygous ATM mutation carriers. *J. Natl Cancer Inst.* 97, 813–822 (2005).
10. Meijers-Heijboer, H. *et al.* Low-penetrance susceptibility to breast cancer due to *CHEK2**1100delC in noncarriers of *BRCA1* or *BRCA2* mutations. *Nature Genet.* 31, 55–59 (2002).
11. Erkko, H. *et al.* A recurrent mutation in *PALB2* in Finnish cancer families. *Nature* 446, 316–319 (2007).
12. Renwick, A. *et al. ATM* mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nature Genet.* 38, 873–875 (2006).
13. Seal, S. *et al.* Truncating mutations in the Fanconi anemia J gene *BRIP1* are low-penetrance breast cancer susceptibility alleles. *Nature Genet.* 38, 1239–1241 (2006).
14. The CHEK2 Breast Cancer Case-Control Consortium. CHEK2*1100delC and susceptibility to breast cancer: A collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from ten studies. *Am. J. Hum. Genet.* 74, 1175–1182 (2004).
15. Cox, A. *et al.* A common coding variant in *CASP8* is associated with breast cancer risk. *Nature Genetics* 39, 352–358 (2007); corrigendum 39, 688 (2007).
16. Easton, D. F. How many more breast cancer predisposition genes are there? *Breast Cancer Res.* 1, 1–4 (1999).
17. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* 27, 234–236 (2001).
18. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* 307, 1072–1079 (2005).
19. Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E. & Begg, C. B. Two-stage designs for gene-disease association studies. *Biometrics* 58, 163–170 (2002).
20. Antoniou, A. C. & Easton, D. F. Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet. Epidemiol.* 25, 190–202 (2003).
21. Altshuler, D. *et al.* A haplotype map of the human genome. *Nature* 437, 1299–1320 (2005).
22. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* 55, 997–1004 (1999).
23. Thomas, D. C., Haile, R. W. & Duggan, D. Recent developments in genomewide association scans: A workshop summary and review. *Am. J. Hum. Genet.* 77, 337–345 (2005).
24. Amundadottir, L. T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nature Genet.* 38, 652–658 (2006).
25. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genet.* 39, 645–649 (2007).
26. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature Genet.* 39, 631–637 (2007).
27. Ferretti, V. *et al.* PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res.* 35, D122–D126 (2007).
28. Moffa, A. B., Tannheimer, S. L. & Ethier, S. P. Transforming potential of alternatively spliced variants of fibroblast growth factor receptor 2 in human mammary epithelial cells. *Mol. Cancer Res.* 2, 643–652 (2004).
29. Adnane, J. *et al.* Bek and Flg, 2 receptors to members of the Fgf family, are amplified in subsets of human breast cancers. *Oncogene* 6, 659–663 (1991).
30. Jang, J. H., Shin, K. H. & Park, J. G. Mutations in fibroblast growth factor receptor 2 and fibroblast growth factor receptor 3 genes associated with human gastric and colorectal cancers. *Cancer Res.* 61, 3541–3543 (2001).
31. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* 446, 153–158 (2007).
32. Lesueur, F. *et al.* Allelic association of the human homologue of the mouse modifier *Ptprj* with breast cancer. *Hum. Mol. Genet.* 14, 2349–2356 (2005).
33. Day, N. *et al.* EPIC-Norfolk: Study design and characteristics of the cohort. *Br. J. Cancer* 80, 95–103 (1999).
34. Breast Cancer Association Consortium. Commonly studied SNPs and breast cancer: Negative results from 12,000 – 32,000 cases and controls from the Breast Cancer Association Consortium. *J. Natl Cancer Inst.* 98, 1382–1396 (2006).
35. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* 30, 38–41 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank the women who took part in this research, and all the funders and support staff who made this study possible. The principal funding for this study was provided by Cancer Research UK. Detailed acknowledgements are provided in Supplementary Information.

Author Contributions D.F.E., A.M.D., P.D.P.P., D.R.C. and B.A.J.P. designed the study and obtained financial support. D.G.B. and D.R.C. directed the genotyping of stages 1 and 2. D.F.E. and D.T. conducted the statistical analysis. K.A.P. and A.M.D. coordinated the genotyping for stage 3 and the fine-scale mapping of the *FGFR2* and *TNRC9* loci. J.P.S. and J.Z. performed resequencing and analysis of the *FGFR2* locus. K.A.P., S.A., C.S.H., R.B., C.A.H., L.K.K., B.E.H., L.L.M., P.B., S.S., V.G., F.O., C-Y. S., P-E.W. and H-C.W. conducted genotyping for the fine-scale mapping. R.L., J.M., H.F. and K.B.M. provided bioinformatics support. D.E., D.G.E., J.P., O.F., N.J., S.S., M.R.S. and N.R. coordinated the studies used in stage 1. N.W. and N.E.D. coordinated the EPIC study used in stages 1 and 2. The remaining authors coordinated the studies in stage 3 and undertook genotyping in those studies. D.F.E. drafted the manuscript, with substantial contributions from K.A.P., A.M.D., P.D.P.P. and B.A.J.P. All authors contributed to the final paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.F.E. (d.easton@srl.cam.ac.uk).

Author affiliations: [1]CR-UK Genetic Epidemiology Unit, Department of Public Health and Primary Care and, [2]Department of Oncology, University of Cambridge, Cambridge CB1 8RN, UK. [3]Perlegen Sciences, Inc., 2021 Stierlin Court, Mountain View, California 94043, USA. [4]Laboratory of Population Genetics, US National Cancer Institute, Bethesda, Maryland 20892, USA. [5]EPIC, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK. [6]MRC Dunn Clinical Nutrition Centre, Cambridge CB2 0XY, UK. [7]Cancer Research UK Cambridge Research Institute, Cambridge CB2 0RE, UK. [8]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA. [9]Epidemiology Program, Cancer Research Center of Hawaii, University of Hawaii, Honolulu, Hawaii 96813, USA. [10]International Agency for Research on Cancer, 150 Cours Albert Thomas, Lyon 69008, France. [11]National Cancer Institute, Bangkok 10400, Thailand. [12]Institute of Biomedical Sciences, Academia Sinica, Taipei 11529, Taiwan. [13]Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton SO16 5YA, UK. [14]Regional Genetic Service, St Mary's Hospital, Manchester M13 0JH, UK. [15]London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK, and Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. [16]Breakthrough Breast Cancer Research Centre, London SW3 6JB, UK. [17]Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. [18]Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [19]Queensland Institute of Medical Research, Brisbane, Queensland 4006, Australia. [20]Departments of Clinical Biochemistry and [21]Breast Surgery, Herlev and Bispebjerg University Hospitals, University of Copenhagen, DK-2730 Herlev, Denmark. [22]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland 20852, USA. [23]Advanced Technology Center, National Cancer Institute, Gaithersburg, Maryland 20877, USA. [24]Cancer Center and M. Sklodowska-Curie Institute of Oncology, Warsaw 02781, Poland. [25]Nofer Institute of Occupational Medicine, Lodz 90950, Poland. [26]Departments of Obstetrics and Gynecology, and [27]Department of Oncology, Helsinki University Central Hospital, Helsinki 00029, Finland. [28]Seoul National University College of Medicine, Seoul 151-742, Korea. [29]National Cancer Center, Goyang 411-769, Korea. [30]Ulsan University College of Medicine, Ulsan 680-749, Korea. [31]Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, 677 Huntington Ave., Boston, Massachusetts 02115, USA. [32]Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Ave., Boston, Massachusetts 02115, USA. [33]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm SE-171 77, Sweden. [34]Population Genetics, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Republic of Singapore. [35]Department of Radiation Oncology and [36]Department of Gynecology and Obstetrics, Hannover Medical School, D-30625 Hannover, Germany. [37]Department of Surgery and [38]Department of Medical Decision Making and [39]Departments of Human Genetics and Pathology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, the Netherlands. [40]Family Cancer Clinic, Department of Medical Oncology, Erasmus MC-Daniel den Hoed Cancer Center, Groene Hilledijk 301, 3075 EA Rotterdam, the Netherlands. [41]Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, Maryland 20892, USA. [42]Environmental Health Sciences, University of Minnesota, Minneapolis, Minnesota 55455, USA. [43]Institute for Cancer Studies and [44]Academic Unit of Surgical Oncology, Sheffield University Medical School, Sheffield S10 2RX, UK. [45]Mayo Clinic College of Medicine, Rochester, Minnesota 55905, USA. [46]VU University Medical Center, 1007 MB Amsterdam, the Netherlands. [47]Department of Clinical Genetics and [48]Internal Medicine, Erasmus University, Rotterdam NL-3015-GE, the Netherlands. [49]Spanish National Cancer Centre (CNIO), Madrid E-28029, Spain. [50]Centre for Molecular, Environmental, Genetic and Analytical Epidemiology, University of Melbourne, Carlton, Victoria 3053, Australia. [51]Department of Preventive and Social Medicine, University of Otago, Dunedin 9001, New Zealand. [52]Cancer Epidemiology Centre, Cancer Council Victoria, Carlton, Victoria 3053, Australia. [53]Genetic Epidemiology

Laboratory, Department of Pathology, University of Melbourne, Parkville, Victoria 3052, Australia. [54]Dr. Margarete Fischer-Bosch-Institute of Clinical Pharamcology, 70376 Stuttgart and University of Tuebingen, 72074 Tuebingen, Germany. [55]Deutsches Krebsforschungszentrum, Heidelberg 69120, Germany. [56]Evangelische Kliniken Bonn gGmbH Johanniter Krankenhaus, 53113 Bonn, Germany. [57]Peter MacCallum Cancer Centre, Melbourne, Victoria 3002, Australia. [58]Insitute of Clinical Medicine, Pathology and Forensic Medicine, University of Kuopio, Kuopio FIN-70210; Finland. [59]Departments of Oncology and Pathology, University Hospital of Kuopio, Kuopio FIN-70211, Finland. [60]Department of Oncology, Vaasa Central Hospital, Vaasa 65130, Finland.

The SEARCH collaborators Craig Luccarini[1], Don Conroy[1], Mitul Shah[1], Hannah Munday[1], Clare Jordan[1], Barbara Perkins[1], Judy West[1], Karen Redman[1] & Kristy Driver[1]. **kConFab** Morteza Aghmesheh[2], David Amor[3], Lesley Andrews[4], Yoland Antill[5], Jane Armes[6], Shane Armitage[7], Leanne Arnold[7], Rosemary Balleine[8], Glenn Begley[9], John Beilby[10], Ian Bennett[11], Barbara Bennett[4], Geoffrey Berry[12], Anneke Blackburn[13], Meagan Brennan[14], Melissa Brown[15], Michael Buckley[16], Jo Burke[17], Phyllis Butow[18], Keith Byron[19], David Callen[20], Ian Campbell[21], Georgia Chenevix-Trench[22], Christine Clarke[23], Alison Colley[24], Dick Cotton[25], Jisheng Cui[26], Bronwyn Culling[27], Margaret Cummings[28], Sarah-Jane Dawson[5], Joanne Dixon[29], Alexander Dobrovic[30], Tracy Dudding[31], Ted Edkins[32], Maurice Eisenbruch[33], Gelareh Farshid[34], Susan Fawcett[35], Michael Field[36], Frank Firgaira[37], Jean Fleming[38], John Forbes[39], Michael Friedlander[40], Clara Gaff[41], Mac Gardner[41], Mike Gattas[42], Peter George[43], Graham Giles[44], Grantley Gill[45], Jack Goldblatt[46], Sian Greening[47], Scott Grist[37], Eric Haan[48], Marion Harris[49], Stewart Hart[50], Nick Hayward[22], John Hopper[51], Evelyn Humphrey[17], Mark Jenkins[52], Alison Jones[7], Rick Kefford[53], Judy Kirk[54], James Kollias[55], Sergey Kovalenko[56], Sunil Lakhani[57], Jennifer Leary[54], Jacqueline Lim[58], Geoff Lindeman[59], Lara Lipton[60], Liz Lobb[61], Mariette Maclurcan[62], Graham Mann[23], Deborah Marsh[63], Margaret McCredie[64], Michael McKay[49], Sue Anne McLachlan[65], Bettina Meiser[4], Roger Milne[26], Gillian Mitchell[49], Beth Newman[66], Imelda O'Loughlin[67], Richard Osborne[51], Lester Peters[68], Kelly Phillips[5], Melanie Price[62], Jeanne Reeve[69], Tony Reeve[70], Robert Richards[71], Gina Rinehart[72], Bridget Robinson[73], Barney Rudzki[74], Elizabeth Salisbury[75], Joe Sambrook[21], Christobel Saunders[76], Clare Scott[5], Elizabeth Scott[77], Rodney Scott[31], Ram Seshadri[37], Andrew Shelling[78], Melissa Southey[26], Amanda Spurdle[22], Graeme Suthers[48], Donna Taylor[79], Christopher Tennant[58], Heather Thorne[21], Sharron Townshend[46], Kathy Tucker[4], Janet Tyler[4], Deon Venter[80], Jane Visvader[81], Ian Walpole[46], Robin Ward[82], Paul Waring[30], Bev Warner[83], Graham Warren[67], Elizabeth Watson[67], Rachael Williams[84], Judy Wilson[85], Ingrid Winship[69] & Mary Ann Young[49]. **AOCS Management Group** David Bowtell[86], Adele Green[22], Anna deFazio[87], Georgia Chenevix-Trench[22], Dorota Gertig[51] & Penny Webb[22].

Consortia affiliations: [1]Department of Oncology, University of Cambridge, Cambridge CB1 8RN, UK. [2]Oncology Research Centre, Prince of Wales Hospital, Randwick, New South Wales 2031, Australia. [3]Genetic Health Services Victoria, Royal Children's Hospital, Melbourne, Victoria 3050, Australia. [4]Hereditary Cancer Clinic, Prince of Wales Hospital, Randwick, New South Wales 2031, Australia. [5]Department of Haematology and Medical Oncology, Peter MacCallum Cancer Centre, St Andrews Place, East Melbourne, Victoria 3002, Australia. [6]Anatomical Pathology, Royal Women's Hospital, Carlton, Victoria 3053, Australia. [7]Molecular Genetics Laboratory, Royal Brisbane and Women's Hospital, Herston, Queensland 4029, Australia. [8]Departments of Translational and Medical Oncology, Westmead Hospital, Westmead, New South Wales 2145, Australia. [9]Cancer Biology Laboratory, TVW Institute for Child Health Research, Subiaco, Western Australia 6008, Australia. [10]Pathology Centre, Queen Elizabeth Medical Centre, Nedlands, Western Australia 6009, Australia. [11]Silverton Place, 101 Wickham Terrace, Brisbane, Queensland 4000, Australia. [12]Department of Public Health and Community Medicine, University of Sydney, Sydney, New South Wales 2006, Australia. [13]John Curtin School of Medical Research, Australian National University, PO Box 334, Canberra, Australian Capital Territory 2601, Australia. [14]NSW Breast Cancer Institute, PO Box 143, Westmead, New South Wales 2145, Australia. [15]Department of Biochemistry, University of Queensland, St. Lucia, Queensland 4072, USA. [16]Molecular and Cytogenetics Unit, Prince of Wales Hospital, Randwick, New South Wales 2031, Australia. [17]Royal Hobart Hospital, GPO Box 1061L, Hobart, Tasmania 7001, Australia. [18]Medical Psychology Unit, Royal Prince Alfred Hospital, Camperdown, New South Wales 2204, Australia. [19]Australian Genome Research Facility, Walter & Eliza Hall Medical Research Institute, Royal Melbourne Hospital, Parkville, Victoria 3050, Australia. [20]Dame Roma Mitchell Cancer Research Laboratories, University of Adelaide/ Hanson Institute, PO Box 14, Rundle Mall, South Australia 5000, Australia. [21]Peter MacCallum Cancer Centre, St Andrew's Place, East Melbourne, Victoria 3002, Australia. [22]Queensland Institute of Medical Research, Herston, Queensland 4006, Australia. [23]Westmead Institute for Cancer Research, University of Sydney, Westmead Hospital ,Westmead, New South Wales 2145, Australia. [24]Department of Clinical Genetics, Liverpool Health Service, PO Box 103, Liverpool, New South Wales 2170, Australia. [25]Mutation Research Centre, St Vincent's Hospital, Victoria Parade, Fitzroy, Victoria 3065, Australia. [26]Centre for Genetic Epidemiology, The University of Melbourne, Level 2 723 Swanston Street, Carlton, Victoria 3053, Australia. [27]Molecular and Clinical Genetics, Level 1 Building 65, Royal Prince Alfred Hospital, Camperdown, New South Wales 2050, Australia. [28]Department of Pathology, University of Queensland Medical School, Herston, New South Wales 4006, Australia. [29]Central Regional Genetic Services, Wellington Hospital, Private bag 7902, Wellington South 6039, New Zealand. [30]Molecular Department of Pathology, Peter MacCallum Cancer Centre, St Andrew's Place, East Melbourne, Victoria 3002, Australia. [31]Hunter Genetics, Hunter Area Health Service, Waratah, New South Wales 2310, Australia. [32]Clinical Chemistry, Princess Margret Hospital for Children, Box D184, Perth, Western Australia 6001, Australia. [33]Department of Multicultural Health, University of Sydney, New South Wales 2052; Australia: [34]Tissue Pathology, Institute of Medical & Veterinary Science, Adelaide, South Australia 5000, Australia. [35]Family Cancer Clinic, Monash Medical Centre, Clayton, Victoria 3168, Australia. [36]Faculty of Medicine, Royal North Shore Hospital, Vindin House, St Leonards, New South Wales 2065, Australia. [37]Department of Haematology, Flinders Medical Centre, Bedford Park, South Australia 5042, Australia. [38]Eskitis Institute of Cell & Molecular Therapies, School of Biomolecular and Biomedical Sciences, Griffith University, Nathan, Queensland 4111, Australia. [39]Surgical Oncology, University of Newcastle, Newcastle Mater Hospital, Waratah, New South Wales 2298, Australia. [40]Department of Medical Oncology, Prince of Wales Hospital, Randwick, New South Wales 2031, Australia.. [41]Victorian Clinical Genetics Service, Royal Melbourne Hospital, Parkville, Victoria 3052, Australia. [42]Queensland Clinical Genetic Service, Royal Children's Hospital, Bramston Terrace, Herston, Queensland 4020, Australia. [43]Clinical Biochemistry Unit, Canterbury Health Labs, PO Box 151, Christchurch 8140, New Zealand. [44]Cancer Epidemiology Centre, The Cancer Council Victoria, 1 Rathdowne Street, Carlton, Victoria 3053, Australia. [45]Department of Surgery, Royal Adelaide Hospital, Adelaide, South Australia 5000, Australia. [46]Genetic Services of WA, King Edward Memorial Hospital, 374 Bagot Road, Subiaco, Western Australia 6008, Australia. [47]Wollongong Hereditary Cancer Clinic, Wollongong Public Hospital, Private Mail Bag 8808, South Coast Mail Centre, New South Wales 2521, Australia. [48]Department of Medical Genetics, Women's and Children's Hospital, North Adelaide, South Australia 5006, Australia. [49]Familial Cancer Clinic, Peter MacCallum Cancer Centre, St Andrew's Place, East Melbourne, Victoria 3002, Australia. [50]Breast and Ovarian Cancer Genetics, Monash Medical Centre, 871 Centre Road, Bentleigh East, Victoria 3165, Australia. [51]Centre for Molecular Environmental, Genetic & Analytic Epidemiology, University of Melbourne, Melbourne, Victoria 3010, Australia. [52]School of Population Health, The University of Melbourne, 723 Swanston Street, Carlton, Victoria 3053, Australia. [53]Medical Oncology, Westmead Hospital, Westmead, New South Wales 2145, Australia. [54]Familial Cancer Service, Department of Medicine, Westmead Hospital, Westmead, New South Wales 2145, Australia. [55]Breast Endocrine and Surgical Unit, Royal Adelaide Hospital, North Terrace, South Australia 5000, Australia. [56]Molecular Pathology Department, Southern Cross Pathology, Monash Medical Centre, Clayton, Victoria 3168, Australia. [57]Molecular and Cellular Pathology, The University of Queensland, Herston, Queensland 4006, Australia. [58]Department of Psychological Medicine, Royal North Shore Hospital, St Leonards, New South Wales 2065, Australia. [59]Breast Cancer Laboratory, Walter and Eliza Hall Institute, PO Royal Melbourne Hospital, Parkville, Victoria 3050, Australia. [60]Medical Oncology and Clinical Haematology Unit, Western Hospital, Footscray, Victoria 3011, Australia. [61]WA Centre for Cancer, Edith Cowan University, Churchlands, Western Australia 6018, Australia. [62]Department of Psychological Medicine, University of Sydney, New South Wales 2006, Australia. [63]Kolling Institute of Medical Research, Royal North Shore Hospital, St Leonards, New South Wales 2065, Australia. [64]Cancer Epidemiology Research Unit, NSW Cancer Council, 153 Dowling Street, Woolloomooloo, New South Wales 2011, Australia. [65]Department of Oncology, St Vincent's Hospital, 41 Victoria Parade, Fitzroy, Victoria 3065, Australia. [66]School of Public Health, Queensland University of Technology, Victoria Park, Kelvin Grove, Queensland 4059, Australia. [67]St Vincent's Breast Clinic, PO Box 4751, Toowoomba, Queensland 4350, Australia. [68]Radiation Oncology, Peter MacCallum Cancer Centre, St Andrew's Place, East Melbourne, Victoria 3002, Australia. [69]Genetic Services, Auckland Hospital, Private Bag 92024, Auckland 1142, New Zealand. [70]Cancer Genetics Laboratory, University of Otago, PO Box 56, Dunedin 9054, New Zealand. [71]Department of Cytogenetics and Molecular Genetics, Women and Children's Hospital, Adelaide, South Australia 5006, Australia. [72]Hancock Family Breast Cancer Foundation, PO Locked Bag 2, West Perth, Western Australia 6005, Australia. [73]Oncology Service, Christchurch Hospital, Private Bag 4710, Christchurch 8140, New Zealand. [74]Molecular Pathology Institute of Medical and Veterinary Science, Frome Road, Adelaide, South Australia 5000, Australia. [75]Section of Cytology, Institute of Clinical Pathology and Medical Research, Westmead Hospital, Westmead, New South Wales 2145, Australia. [76]School of Surgery and Pathology, QE11 Medical Centre, M block 2nd Floor, Nedlands, Western Australia 6907, Australia. [77]South View Clinic, Suite 13, Level 3 South Street, Kogarah, New South Wales 2217, Australia. [78]Department of Obstetrics and Gynaecology, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand. [79]Department of Radiology, Royal Perth Hospital, Box X2213, Perth 6011, Western Australia, Australia. [80]Murdoch Institute, Royal Children's Hospital, Parkville, Victoria 3050, Australia. [81]Molecular Genetics of Cancer Division, Walter & Eliza Hall Medical Research Institute, Royal Melbourne Hospital, Parkville, Victoria 3050, Australia. [82]Department of Medical Oncology, St Vincents Hospital, Darlinghurst, New South Wales 2010, Australia. [83]Cabrini Hospital, 183 Wattletree Road, Malvern, Victoria 3144, Australia. [84]Family Cancer Clinic, St Vincent's Hospital, Darlinghurst, New South Wales 2010, Australia. [85]Medical Psychology Research Unit, Royal North Shore Hospital, St Leonards, New South Wales 2065, Australia. [86]Cancer Genomics & Biochemistry Laboratory, Peter MacCallum Cancer Centre, St Andrew's Place, East Melbourne, Victoria 3002, Australia. [87]Obstetrics & Gynaecology, Westmead Hospital, University of Sydney, New South Wales 2006, Australia.

METHODS

Subjects. Cases in stage 1 were identified through clinical genetics centres in Cambridge ($n = 91$), Manchester (96) and Southampton (136), and a national study of bilateral breast cancer (85). Cases were women diagnosed with invasive breast cancer under the age of 60 years who had a family history score of at least 2, where the score was computed as the total number of first-degree relatives plus half the number of second-degree relatives affected with breast cancer. The score for women with bilateral breast cancer was increased by 1, so that women were eligible if they were diagnosed with bilateral breast cancer and had one affected first-degree relative. Cases known to carry a *BRCA1* or *BRCA2* mutation were excluded. Controls were selected from the EPIC-Norfolk study, a population-based cohort study of diet and cancer based in Norfolk, East Anglia, UK[33]. Controls were chosen to be women aged over 50 years and free of cancer at the time of entry. Genotyping was attempted on 408 cases, plus 32 duplicate case samples, and 400 controls. For the analysis in Table 1, 54 samples with genotype call rates <80% were excluded, so the final analyses were based on 390 cases and 364 controls. The minimum genotype call rate for the remaining samples was 89%. The overall genotype discordance rate between duplicate samples in stage 1 was 0.01%.

For stage 2, invasive breast cancer cases were drawn from SEARCH, a population-based study of cancer in East Anglia[32]. Controls were women selected from the EPIC-Norfolk study, as previously described[33]. Eighty-eight subjects who were also genotyped in stage 1, and 35 controls who subsequently developed breast cancer and were also in the case series, were excluded from the analysis, leaving 3,990 breast cancer cases and 3,916 controls, plus five duplicates. The overall rate of discordance of genotypes between duplicate samples in stage 2 was 0.008%.

Twenty-one additional studies were included in stage 3 (see Supplementary Table 2). These studies participated through the Breast Cancer Association Consortium, an ongoing collaboration among investigators conducting case-control association studies in breast cancer[15,33]. All studies provided information on disease status (invasive breast cancer, carcinoma *in situ* or control), age at diagnosis/observation, ethnic group, first-degree family history of breast cancer and bilaterality of breast cancer. One further study (Breast Cancer Study of Taiwan) was included in the fine-scale mapping of the *FGFR2* locus.

Genotyping. For stage 1, genotyping was performed on 200 ng DNA that was first subjected to whole genome amplification using Multiple Displacement Amplification (MDA)[36]. Samples were then genotyped for a set of 266,732 SNPs using high-density oligonucleotide, photolithographic microarrays at Perlegen Sciences. For stage 2, genotyping was performed using 2.5 μg genomic DNA. These samples were genotyped for a set of 13,023 SNPs selected on the basis of the stage 1 results, using a custom designed oligonucleotide array. For both stages, each SNP was interrogated by 24 25-mer oligonucleotide probes synthesized by photolithography on a glass substrate. The 24 features comprise 4 sets of 6 features interrogating the neighbourhoods of SNP reference and alternative alleles on forward and reference strands. Each allele and strand is represented by five offsets: -2, -1, 0, 1 and 2 indicating the position of the SNP within the 25-mer, with zero being at the thirteenth base. At offset 0 a quartet was tiled, which included the perfect match to reference and alternative SNP alleles, and the two remaining nucleotides as mismatch probes. When possible, the mismatch features were selected as a purine nucleotide substitution for a purine perfect match nucleotide and a pyrimidine nucleotide substitution for a pyrimidine perfect match nucleotide. Thus, each strand and allele tiling consisted of 6 features comprising five perfect match probes and one mismatch.

Individual genotypes were determined by clustering all SNP scans in the two-dimensional space defined by reference and alternative trimmed mean intensities, corrected for background. Allele frequencies were approximated using the intensities collected from the high-density oligonucleotide arrays. An SNP's allele frequency, p, was estimated as the ratio of the relative amount of the DNA with reference allele to the total amount of DNA. The $\hat{p}$ value was computed from the trimmed mean intensities of perfect match features, after subtracting a measure of background computed from trimmed means of intensities of mismatch features. The trimmed mean disregarded the highest and the lowest intensity from the five perfect match intensities before computing the arithmetic mean. For the mismatch features, the trimmed mean is the individual intensity of the specified mismatch feature.

The genotype clustering procedure was an iterative algorithm developed as a combination of K-means and constrained multiple linear regressions. The K-means at each step re-evaluated the cluster membership representing distinct diploid genotypes. The multiple linear regressions minimized the variance in $\hat{p}$ within each cluster while optimizing the regression lines' common intersect. The common intersect defined a measure of common background that was used to adjust the allele frequencies for the next step of K-means. The K-means and multiple linear regression steps were iterated until the cluster membership and background estimates converged. The best number of clusters was selected by maximizing the total likelihood over the possible cluster counts of 1, 2 and 3 (representing the combinations of the three possible diploid genotypes). The total likelihood was composed of data likelihood and model likelihood. The data likelihood was determined using a normal mixture model for the distribution of $\hat{p}$ around the cluster means. The model likelihood was calculated using a prior distribution of expected cluster positions, resulting in optimal $\hat{p}$ positions of 0.8 for the homozygous reference cluster, 0.5 for the heterozygous cluster and 0.2 for the homozygous alternative cluster.

A genotyping quality metric was compiled for each genotype from 15 input metrics that described the quality of the SNP and the genotype. The genotyping quality metric correlated with a probability of having a discordant call between the Perlegen platform and outside genotyping platforms (that is, non-Perlegen HapMap project genotypes). A system of 10 bootstrap aggregated regression trees was trained using an independent data set of concordance data between Perlegen genotypes and HapMap project genotypes. The trained predictor was then used to predict the genotyping quality for each of the genotypes in this data set. Genotypes with quality scores of less than 7 were discarded. Data were analysed for 227,876 SNPs in stage 1 and 12,026 (of 13,023 selected) in stage 2, for which the call rate was >80%.

The 12,711 SNPs for stage 2 were primarily selected on the basis of a 1 d.f. Cochran-Armitage trend test (11,809, all with $P < 0.052$). We also included 826 SNPs with $P < 0.01$ testing for the difference in frequency of either homozygote between cases and controls (that is, assuming either a dominant or recessive model) and 76 SNPs that achieved $P < 0.01$ on a Cochran-Armitage test, weighting individuals by their family history score as above.

For the main analyses, we discarded SNPs with a call rate <90% in stage 1 and 95% in stage 2, and SNPs with a deviation from Hardy–Weinberg equilibrium significant at $P < 0.00001$ in either stage, leaving 205,586 SNPs in stage 1 and 10,621 SNPs in stage 2.

The 30 SNPs included in the stage 3 analyses were initially selected on the basis of a combined analysis of stage 1 and stage 2. We included all SNPs achieving a combined $P < 0.00002$ (based on either the Cochran-Armitage or 2 d.f. test, see below). Following re-evaluation of the stage 2 genotyping by 5′ nuclease assay (Taqman, Applied Biosystems) using the ABI PRISM 7900HT (Applied Biosystems), and exclusion of some samples, 16 of these SNPs were significant at $P < 0.00002$ and 24 at $P < 0.0002$ (Supplementary Table 3). One additional SNP, rs3803662, was added as a result of fine-scale mapping of the *TNRC9/LOC643714* locus.

The 31 stage 3 SNPs were genotyped in 22 studies (including cases and controls from SEARCH not used in stage 2, together with 21 other studies). For 18 of the studies, genotyping was performed by 5′ nuclease assay (Taqman) using the ABI PRISM 7900HT or 7500 Sequence Detection Systems according to manufacturer's instructions. Primers and probes were supplied directly by Applied Biosystems (http://www.appliedbiosystems.com/) as Assays-by-Design. All assays were carried out in 384-well or 96-well format, with each plate including negative controls (with no DNA). Duplicate genotypes were provided for at least 2% of samples in each study. For three studies, SNPs were genotyped using matrix assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) for the determination of allele-specific primer extension products using Sequenom's MassARRAY system and iPLEX technology. The design of oligonucleotides was carried out according to the guidelines of Sequenom and performed using MassARRAY Assay Design software (version 1.0). Multiplex PCR amplification of amplicons containing SNPs of interest was performed using Qiagen HotStart Taq Polymerase on a Perkin Elmer GeneAmp 2400 thermal cycler (MJ Research) with 5 ng genomic DNA. Primer extension reactions were carried out according to manufacturer's instructions for iPLEX chemistry. Assay data were analysed using Sequenom TYPER software (version 3.0). One study used both the Taqman and MALDI-TOF MS approaches. The SNPs genotyped in stage 3 were also regenotyped in the stage 2 samples using Taqman; these genotype calls were used in the overall analyses (Table 2, Supplementary Table 3, and Fig. 2).

We eliminated any sample that could not be scored on 20% of the SNPs attempted. We also removed data for any centre/SNP combination for which the call rate was less than 90%. In any instances where the call rate was 90–95%, the clustering of genotype calls was re-evaluated by an independent observer to determine whether the clustering was sufficiently clear for inclusion. We also eliminated all the data for a given SNP/centre where the reproducibility in duplicate samples was <97%, or where there was marked deviation from Hardy–Weinberg equilibrium in the controls ($P < 0.00001$).

Fine-scale mapping of *FGFR2*. Initial tagging of the associated region was done by identifying all SNPs with an m.a.f. > 5% in the HapMap CEPH/CEU set (Utah residents with ancestry from northern and western Europe). We then selected 7 SNPs (in addition to rs2981582) that tagged these variants with a pairwise $r^2 > 0.8$, using the program Tagger (http://www.broad.mit.edu/mpg/tagger/)[37]. To identify additional common variants within the 32.5 kb region of linkage around the associated SNP, we resequenced 45 lymphocyte DNA samples from a subset of European subjects also genotyped by HapMap and other publicly available data sets. Seventy overlapping PCR amplicons were designed from positions 123317613 to 123348192 of chromosome 10 (average amplicon size 650 bp, 160 bp overlap). M13-tagged PCR products were bidirectionally sequenced using Big Dye 3.0 (Applied Biosystems) and processed using automated trace analysis through the Cancer Genome Workbench (cgwb.nci.nih.gov). Eighty-six per cent of the nucleotides across the region could be scored for polymorphisms in at least 80% of subjects. This set gave a >97% probability of detecting a variant with an m.a.f. > 5%. One hundred and seventeen variants were identified, including 27 present in dbSNP but without individual genotype information in European subjects, and an additional 46 not in dbSNP. Individual genotype information was then compared and merged with publicly available genotypes from Caucasian subjects (HapMap release 21 for 60 CEU parents, 22 European subjects from the Environmental Genome Project (EGP) resequencing effort (http://egp.gs.washington.edu/data/fgfr2/), and 24 European subjects from Perlegen (retrieved through http://gvs.gs.washington.edu/GVS)). There were 2 discrepancies among 389 genotype calls among subjects in common between our resequencing effort and EGP or Perlegen data, and 10 out of 926 compared to HapMap genotypes.

On the basis of these data, we identified 28 SNPs correlated with rs2981582 with $r^2 > 0.6$. We then attempted to genotype these 28 SNPs, plus rs2981582, in a subset of 80 controls from SEARCH and 84 controls from the Seoul Breast Cancer Study. Twenty-two of the variants were genotyped using Taqman. Four further variants (rs34032268, rs2912778, rs2912781 and rs7895676), which were not amenable to Taqman, were genotyped by Pyrosequencing (Biotage; http://www.biotagebio.com/). Assays were designed using Pyrosequencing Assay Design Software 1.0. The remaining 2 SNPs (rs35393331 and rs33971856) could not be genotyped using either technology and were excluded from further analyses. We cannot therefore comment on their likelihood of being the causal variant. Using these data, we selected tagging sets of 11 SNPs for UK subjects and 14 SNPs for Korean subjects (including rs2981582), such that each of the remaining variants was correlated with a tagging SNP with $r^2 > 0.95$ in the UK study or $r^2 > 0.86$ in the Korean study. After genotyping the 11 tag SNPs in SEARCH, two of these SNPs (rs4752569 and rs35012336) showed strong evidence against being the causative variant and were not considered further. The remaining 12 tag SNPs from the Korean subset were then genotyped in the samples from the IARC-Thai Breast Cancer Study, the Breast Cancer Study in Taiwan and the Multi-Ethnic Cohort (MEC), by Taqman.

Statistical methods. The primary test used for each SNP was a Cochran-Armitage 1 d.f. score test for association between disease status and allele dose. In the combined analysis, we performed a stratified Cochran-Armitage test. Stage 1 was given a weight of 4 in this analysis (corresponding to a weight of 2 in the score statistic), to allow for the expected greater effect size given the inclusion of cases with a family history. In the stage 3 analyses, each study was treated as a separate stratum, except for the MEC, in which the European American and Japanese American subgroups were treated as separate strata. For all studies except the MEC, individuals from a minor ethnic group for that study were excluded. Per-allele and genotype-specific odds ratios, and confidence intervals, were estimated using logistic regression, adjusting for the same strata. The summary odds ratios in Fig. 2 are based on the data from the stage 3 studies only, to avoid the bias inherent in estimates from the stage 1 and 2 data for SNPs exhibiting an association (the so called 'winner's curse'). The effects of genotype on family history of breast cancer (first degree yes/no) and bilaterality were examined by treating these variables as outcomes in a stratified Cochran-Armitage test.

To assess the global significance of the SNPs in stage 3, we computed the sum of the χ^2 trend statistics (excluding the 6 SNPs reaching genome-wide significance, plus rs2107425 as it was in LD with rs3817198) over those SNPs (17 of 23) for which the estimated odds ratios in stage 3 were in the same direction as the combined stage 1/stage 2[38]. Under the null hypothesis of no association, the asymptotic distribution of this statistic is χ^2 with n degrees of freedom, where n has a binomial distribution with parameters 23 and 1/2. The significance of this statistic was then assessed by computing a weighted sum of the tails of the relevant χ^2 distributions.

For the fine-scale mapping of the *FGFR2* locus, we first derived haplotype frequencies using the haplo.stats package in S-plus[39], separately for the European and Asian populations, using data from the case-control studies on whom the tag SNPs were typed plus the 164 control individuals on whom all SNPs were typed. These were used to impute genotype probabilities for each identified SNP in each individual. We then used an EM algorithm to fit a logistic regression model assuming that each SNP in turn was the causal variant, allowing for uncertainty in the genotypes of untyped SNPs, and hence to determine the likelihood that each SNP was the causal variant.

Coverage of the stage 1 tagging set was estimated using HapMap phase II as a reference. We based estimates on 2,116,183 SNPs with an m.a.f. of >5% in the CEU population. Of the SNPs successfully genotyped in stage 1, 187,663 were also on HapMap. For those SNPs not on HapMap, we identified 'surrogate' SNPs that were in perfect LD based on genotyping of 24 Caucasians by Perlegen Sciences (269,203 SNPs)[18]. To estimate coverage, we determined the best pairwise r^2 for each HapMap SNP and each tag SNP or a surrogate SNP, using the HapMap CEU data. This coverage was summarized in terms of the distribution of r^2 by allele frequency in 10 categories.

To estimate the power to detect each of the associations found, we computed the non-centrality parameter for the test statistic at each stage, based on the per-allele relative risk, allele frequency and r^2. This was used to estimate the power for a given r^2, based on a simulated trivariate normal distribution for the score statistics after each stage to allow for the correlations in the test statistics. We assumed a cut-off of $P < 0.05$ for stage 1, $P < 0.00002$ for stage 2 and $P < 10^{-7}$ for stage 3 (the first is slightly conservative, as more SNPs than this were actually taken forward). The overall power was obtained by averaging the power estimates for each r^2 over the distribution of r^2 obtained from the HapMap data, applicable to a SNP of that frequency.

The expected number of significant associations after stage 2 (Table 1) was calculated using a bivariate normal distribution for the joint distribution of the (weighted) Cochran-Armitage score statistics after stage 1 and after both stages, using a correlation of 0.525 between the two statistics (reflecting the weighted sizes of the two studies). These calculations were based on the 205,586 SNPs reaching the required quality control in stage 1. Of these, 11,313 reached a $P < 0.05$, of which 7,405 (65.5%) were successfully genotyped to the required quality control in stage 2. Thus the expected number reaching a given significance level with good quality control was calculated from the total number expected to reach this level × 65.5%. We adjusted the variances of the test statistics, separately for stages 1 and 2, using the genomic control method[22]. The adjustment factor, λ, was estimated from the median of the smallest 90% of the test statistics for SNPs typed in that stage, divided by the predicted median for the smallest 90% of a sample of χ^2_1 distributions (that is, the 45% percentile of a χ^2_1 distribution, 0.375).

[36] Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA* 99, 5261–5266 (2002).

[37] de Bakker, P. I. W. *et al.* Efficiency and power in genetic association studies. *Nature Genet.* 37, 1217–1223 (2005).

[38] Tyrer, J., Pharoah, P. D. P. & Easton, D. F. The admixture maximum likelihood test: A novel experiment-wise test of association between disease and multiple SNPs. *Genet. Epidemiol.* 30, 636–643 (2006).

[39] Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. & Poland, G. A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* 70, 425–434 (2002).

nature genetics

Two variants on chromosome 17 confer prostate cancer risk, and the one in *TCF2* protects against type 2 diabetes

Julius Gudmundsson[1,30], Patrick Sulem[1,30], Valgerdur Steinthorsdottir[1], Jon T Bergthorsson[1], Gudmar Thorleifsson[1], Andrei Manolescu[1], Thorunn Rafnar[1], Daniel Gudbjartsson[1], Bjarni A Agnarsson[2], Adam Baker[1], Asgeir Sigurdsson[1], Kristrun R Benediktsdottir[2], Margret Jakobsdottir[1], Thorarinn Blondal[1], Simon N Stacey[1], Agnar Helgason[1], Steinunn Gunnarsdottir[1], Adalheidur Olafsdottir[1], Kari T Kristinsson[1], Birgitta Birgisdottir[1], Shyamali Ghosh[1], Steinunn Thorlacius[1], Dana Magnusdottir[1], Gerdur Stefansdottir[1], Kristleifur Kristjansson[1], Yu Bagger[3], Robert L Wilensky[4], Muredach P Reilly[4], Andrew D Morris[5], Charlotte H Kimber[6], Adebowale Adeyemo[7], Yuanxiu Chen[7], Jie Zhou[7], Wing-Yee So[8], Peter C Y Tong[8], Maggie C Y Ng[8], Torben Hansen[9], Gitte Andersen[9], Knut Borch-Johnsen[9–11], Torben Jorgensen[11], Alejandro Tres[12,13], Fernando Fuertes[14], Manuel Ruiz-Echarri[12], Laura Asin[13], Berta Saez[13], Erica van Boven[15], Siem Klaver[16], Dorine W Swinkels[16], Katja K Aben[17], Theresa Graif[18], John Cashy[18], Brian K Suarez[19], Onco van Vierssen Trip[20], Michael L Frigge[1], Carole Ober[21], Marten H Hofker[22,23], Cisca Wijmenga[24,25], Claus Christiansen[3], Daniel J Rader[4], Colin N A Palmer[6], Charles Rotimi[7], Juliana C N Chan[8], Oluf Pedersen[9,10], Gunnar Sigurdsson[26,27], Rafn Benediktsson[26,27], Eirikur Jonsson[28], Gudmundur V Einarsson[28], Jose I Mayordomo[12,13], William J Catalona[18], Lambertus A Kiemeney[29], Rosa B Barkardottir[2], Jeffrey R Gulcher[1], Unnur Thorsteinsdottir[1], Augustine Kong[1] & Kari Stefansson[1]

We performed a genome-wide association scan to search for sequence variants conferring risk of prostate cancer using 1,501 Icelandic men with prostate cancer and 11,290 controls. Follow-up studies involving three additional case-control groups replicated an association of two variants on chromosome 17 with the disease. These two variants, 33 Mb apart, fall within a region previously implicated by family-based linkage studies on prostate cancer. The risks conferred by these variants are moderate individually (allele odds ratio of about 1.20), but because they are common, their joint population attributable risk is substantial. One of the variants is in *TCF2* (*HNF1β*), a gene known to be mutated in individuals with maturity-onset diabetes of the young type 5. Results from eight case-control groups, including one West African and one Chinese, demonstrate that this variant confers protection against type 2 diabetes.

[1]deCODE genetics, Sturlugata 8, 101 Reykjavik, Iceland. [2]Department of Pathology, Landspitali-University Hospital, 101 Reykjavik, Iceland. [3]Center for Clinical and Basic Research A/S, DK-2750 Ballerup, Denmark. [4]University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA. [5]Division of Medicine and Therapeutics, Ninewells Hospital and Medical School, Dundee DD1 9SY, Scotland. [6]Population Pharmacogenetics Group, Biomedical Research Centre, Ninewells Hospital and Medical School, Dundee DD1 9SY, Scotland. [7]National Human Genome Center, Howard University, Department of Community and Family Medicine, Washington, DC 20060, USA. [8]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, NT, Hong Kong. [9]Steno Diabetes Center, DK-2820 Copenhagen, Denmark. [10]Faculty of Health Science, University of Aarhus, DK-8000 Aarhus, Denmark. [11]Research Centre for Prevention and Health, Glostrup University Hospital, DK-2600 Glostrup, Denmark. [12]Division of Medical Oncology, Lozano Blesa University Hospital, University of Zaragoza, 50009 Zaragoza, Spain. [13]The Institute of Health Sciences, Nanotechnology Institute of Aragon, 50009 Zaragoza, Spain. [14]Division of Radiation Oncology, Lozano Blesa University Hospital, University of Zaragoza, 50009 Zaragoza, Spain. [15]Department of Urology, Maas Ziekenhuis, 5830 AB Boxmeer, The Netherlands. [16]Department of Clinical Chemistry, Radboud University Nijmegen Medical Center, 6500 HB Nijmegen, The Netherlands. [17]Comprehensive Cancer Center East, 6501 BG Nijmegen, and Department of Epidemiology and Biostatistics, Radboud University Nijmegen Medical Center, 6500 HB Nijmegen, The Netherlands. [18]Department of Urology, Northwestern University Feinberg School of Medicine, Chicago, Illinois 60611, USA. [19]Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri 63110, USA. [20]Department of Urology, Gelderse Vallei Hospital, 6716 RP Ede, The Netherlands. [21]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. [22]Department of Pathology and Laboratory Medicine, University Medical Center Groningen, 9700 RB Groningen, The Netherlands. [23]Department of Molecular Genetics, Maastricht University, 6200 MD Maastricht, The Netherlands. [24]Department of Genetics, University Medical Center Groningen, 9700 RB Groningen, The Netherlands. [25]Complex Genetics Section, Department of Biomedical Genetics, University Medical Centre Utrecht, 3508 AB Utrecht, The Netherlands. [26]Landspitali-University Hospital, 101 Reykjavik, Iceland. [27]Icelandic Heart Association, 201 Kopavogur, Iceland. [28]Department of Urology, Landspitali-University Hospital, 101 Reykjavik, Iceland. [29]Department of Epidemiology and Biostatistics and Department of Urology, Radboud University Nijmegen Medical Center, 6500 HB Nijmegen, The Netherlands. [30]These authors contributed equally to this work. Correspondence should be addressed to K.S. (kstefans@decode.is) or J.G. (julius.gudmundsson@decode.is).

Received 21 March; accepted 8 May; published online 1 July 2007; doi:10.1038/ng2062

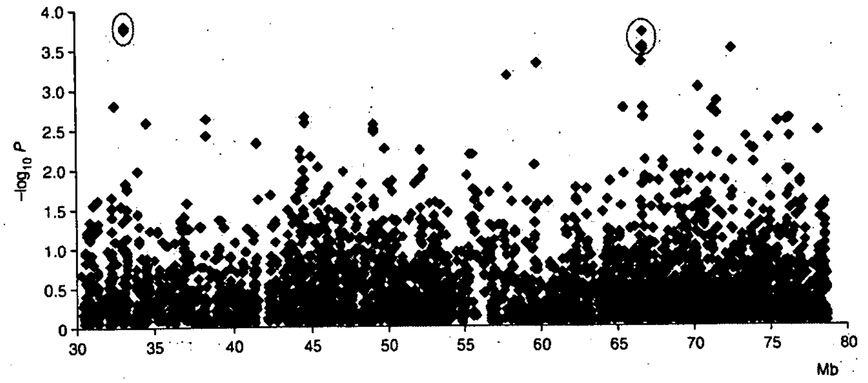


Figure 1 A schematic view of the genome-wide association results for chromosome 17q. Shown are results from the genome-wide association analysis performed in the Icelandic study population. The results plotted are for all Illumina Hap300 chip SNPs that are located between position 30 Mb and the telomere (~78.6 Mb; build 35) on the long arm of chromosome 17 (blue diamonds). The six SNP markers circled in red and listed in **Table 2** all fall within the linkage region described in ref. 8.

Firmly established risk factors for prostate cancer are age, ethnicity and family history. Despite a large body of evidence for a genetic component to the risk of prostate cancer, sequence variants on 8q24 are the only common variants reported so far that account for substantial proportion of cases[1–4].

In the present study, we began with a genome-wide SNP association study, applying 310,520 SNPs from the Illumina Hap300 chip to search for sequence variants conferring risk of prostate cancer using Icelandic cases and controls. We expanded the data from a previously reported study[2] by increasing the number of cases from 1,453 to 1,501 and the number of controls from 3,064 to 11,290. This corresponds to a ~34% increase in effective sample size. Apart from the variants on 8q24 (refs. 1,2) and SNPs correlated with them, no other SNPs achieved genome-wide significance (**Supplementary Fig. 1** online). However, we assumed that a properly designed follow-up strategy would lead to the identification of additional susceptibility variants for prostate cancer.

Like others[5], we believe that results from family-based linkage studies should be taken into account when evaluating the association results of a genome-wide study. However, instead of using linkage scores to formally weight the statistical significance of different SNPs[5], we used them to prioritize follow-up studies. The long arm of chromosome 17 has been reported in several linkage studies of prostate cancer[6–8], but no susceptibility variants have yet been found[9–11]. Hence, we decided to focus on this region first.

We selected for further analysis six SNPs on chromosome 17q having the lowest P values ($<5 \times 10^{-4}$) and ranking from 68 to 100 among the most significantly associated SNPs in our genome-wide analysis (**Fig. 1**). These SNPs mapped to two distinct regions on chromosome 17q that are both within a region with LOD scores ranging from 1–2 but outside the proposed 10-cM candidate gene region reported in a recent linkage analysis[8]. One locus was on 17q12 (rs7501939 and rs3760511), encompassing the 5′ end of the *TCF2* (*HNF1β*) gene, where the linkage disequilibrium (LD) is weak (based on the Utah CEPH (CEU) HapMap data set). The second locus is in a gene-poor area on 17q24.3 (rs1859962, rs7214479, rs6501455 and rs983085) where all four SNPs fall within a strong LD block (based on the CEU HapMap data set). The two loci are separated by approximately 33 Mb, and we did not observe any LD between them (see **Supplementary Table 1** online for r^2 and D' values).

We genotyped five of the six SNPs in three prostate cancer case-control groups of European ancestry (**Table 1**). The assay for rs983085 on 17q24.3 failed in genotyping, but this SNP is almost perfectly correlated with rs6501455 ($r^2 = 0.99$) and is therefore expected to give comparable results. For each of the replication study groups, the observed effect of four of the five SNPs were in the same direction as in Iceland. One SNP, rs6501455, showed an opposite effect in the Chicago group. When results from all four case-control groups were combined, two SNPs achieved genome-wide significance, rs7501939 allele C (rs7501939 C) at 17q12 (odds ratio (OR) = 1.19, $P = 4.7 \times 10^{-9}$) and rs1859962 allele G (rs1859962 G) at 17q24.3 (OR = 1.20, $P = 2.5 \times 10^{-10}$) (**Tables 2 and 3**). In an effort to refine the signal at the 17q12 locus, we selected three SNPs (rs4239217, rs757210, rs4430796) that were substantially correlated with rs7501939 ($r^2 > 0.5$) based on the CEU HapMap data. One of these, rs4430796, showed an association to prostate cancer that was stronger than that of rs7501939. Specifically, with all groups combined, allele A of rs4430796 had an OR of 1.22 with a P of 1.4×10^{-11} (**Table 2**). A joint analysis showed that the effects of rs7501939 and rs3760511 were no longer significant after adjusting for rs4430796 ($P = 0.88$ and 0.58, respectively), whereas rs4430796 remained significant after adjusting for both rs7501939 and rs3760511 ($P = 0.0042$). At 17q24.3, our attempt at refining the signal did not result in any SNP that was more significant than rs1859962. Among the Illumina SNPs, rs7114479 and rs6501455 were not significant ($P > 0.75$) with adjustment for the effect of rs1859962, whereas rs1859962 remained significant after adjusting for the other two SNPs ($P = 7.4 \times 10^{-4}$). Henceforth, our focus was on rs4430796 at 17q12 and rs1859962 at 17q24.3. However, at 17q12, because rs7501539 was a part of the original genome-wide scan, we have included it in the discussion when appropriate. For replication efforts, we recommend including at least the three abovementioned SNPs. We note that in the results released by the Cancer Genetic Markers of Susceptibility study group (see URL below), these three SNPs also show nominal, but not genome-wide, significant association with prostate cancer.

For men with prostate cancer diagnosed at age 65 or younger, the observed OR from the combined analysis was slightly higher (1.30

Table 1 Characteristics of men with prostate cancer and controls from four sources

Study population	Affected individuals	Controls	Aggressive[a] (%)	Mean age at diagnosis (range)	Age at diagnosis <65 years (%)
Iceland	1,501	11,290	50	70.8 (40–96)	22
Nijmegen, The Netherlands	999	1,466	47	64.2 (43–83)	52
Zaragoza, Spain	456	1,078	37	69.3 (44–83)	19
Chicago	537	514	48	59.6 (39–87)	70
Total:	3,493	14,348			

[a] 'Aggressive' is defined here as cancers with Gleason scores of 7 or higher and/or a stage of T3 or higher and/or node-positive disease and/or metastatic disease.

Table 2 Association results for SNPs on 17q12 and prostate cancer in Iceland, The Netherlands, Spain and the US

	Frequency			
Study population (*N* cases/*N* controls) and variant (allele)	Cases	Controls	OR (95% c.i.)	*P* value
Iceland (1,501/11,289)				
rs7501939 (C)	0.615	0.578	1.17 (1.08–1.27)	1.8×10^{-4}
rs3760511 (C)	0.384	0.348	1.17 (1.08–1.27)	1.6×10^{-4}
rs4430796 (A)	0.558	0.512	1.20 (1.11–1.31)	1.4×10^{-5}
The Netherlands (997/1,464)				
rs7501939 (C)	0.648	0.589	1.29 (1.15–1.45)	2.4×10^{-5}
rs3760511 (C)	0.362	0.338	1.11 (0.99–1.25)	0.086
rs4430796 (A)	0.568	0.508	1.28 (1.14–1.43)	3.1×10^{-5}
Spain (456/1,078)				
rs7501939 (C)	0.583	0.566	1.07 (0.92–1.26)	0.37
rs3760511 (C)	0.277	0.257	1.11 (0.93–1.32)	0.25
rs4430796 (A)	0.469	0.454	1.06 (0.91–1.24)	0.45
Chicago (536/514)				
rs7501939 (C)	0.637	0.588	1.15 (1.03–1.47)	0.021
rs3760511 (C)	0.347	0.294	1.28 (1.06–1.54)	9.4×10^{-3}
rs4430796 (A)	0.563	0.477	1.41 (1.19–1.67)	9.4×10^{-5}
All excluding Iceland (1,989/3,056)[a]				
rs7501939 (C)	–	0.581	1.21 (1.12–1.32)	5.6×10^{-6}
rs3760511 (C)	–	0.296	1.15 (1.05–1.25)	2.4×10^{-3}
rs4430796 (A)	–	0.480	1.24 (1.14–1.35)	2.0×10^{-7}
All combined (3,490/14,345)[a]				
rs7501939 (C)	–	0.580	1.19 (1.12–1.26)	4.7×10^{-9}
rs3760511 (C)	–	0.309	1.16 (1.09–1.23)	1.4×10^{-6}
rs4430796 (A)	–	0.488	1.22 (1.15–1.30)	1.4×10^{-11}

All P values shown are two sided. Shown are the numbers of cases and controls (*N*), allelic frequencies of variants in affected and control individuals, the allelic odds ratio (OR) with 95% confidence interval (95% c.i.) and *P* values based on the multiplicative model.

[a] For the combined study populations, the reported control frequency was the average, unweighted control frequency of the individual populations, whereas the OR and the *P* values were estimated using the Mantel-Haenszel model.

for rs4430796 A and 1.27 for rs1859962 G). For each copy of the at-risk alleles, carriers were diagnosed with prostate cancer 2 months younger for rs4430796 and 5 months younger for rs1859962, compared with noncarriers with prostate cancer. However, this observation was not statistically significant and therefore requires further investigation.

We did not observe any interaction between the risk variants on 17q12 and 17q24.3; a multiplicative or log-additive model provided an adequate fit for the joint risk of rs4430796 and rs1859962. We estimated genotype-specific ORs for each locus individually (**Table 4**). Based on results from all four groups, a multiplicative model for the genotype risk provided an adequate fit for rs4430796 at 17q12. However, for rs1859962 at the 17q24.3 locus, the full model provided a significantly better fit than the multiplicative model ($P = 0.006$), a result driven mainly by the Icelandic samples. Specifically, the estimated OR of 1.33 for a heterozygous carrier of rs1859962 G was substantially higher than the 1.20 estimate implied by a multiplicative model.

The SNPs rs7501939 and rs4430796 on 17q12 are located in the first and second intron of the *TCF2* gene, respectively. To the best of our knowledge, sequence variants in *TCF2* have not been previously implicated in the risk of prostate cancer. More than 50 different exonic *TCF2* mutations have been reported in individuals with renal cysts, maturity-onset diabetes of the young type 5 (MODY5), pancreatic atrophy and genital tract abnormalities[12,13]. We sequenced all nine exons of *TCF2* in 200 Icelandic men with prostate cancer and 200 Icelandic controls without detecting any mutations explaining our association signal (data not shown).

Notably, several epidemiological studies have demonstrated an inverse relationship between type 2 diabetes (T2D) and the risk of prostate cancer (see ref. 14 and references therein). A recent meta-analysis estimated the relative risk of prostate cancer to be 0.84 (95% confidence interval (c.i.), 0.71–0.92) among diabetes patients[14]. Therefore, we decided to investigate a potential association between T2D and the SNPs in *TCF2* showing the strongest association with prostate cancer in our data.

We typed the Illumina SNP rs7501939 in 1,380 individuals with T2D (males in this group were not known to have prostate cancer, according to the Icelandic Cancer Registry list of individuals with prostate cancer diagnosed from 1955 to 2006). When compared with 9,940 controls not known to have either prostate cancer or T2D, rs7501939 C showed a protective effect against T2D (OR = 0.88, $P = 0.0045$) in these samples. For the same samples, allele A of the refinement SNP rs4430796 gave a comparable result (OR = 0.86, $P = 0.0021$). To validate this association, we typed both rs7501939 and rs4430796 in seven additional T2D case-control groups of European, African and Asian ancestry (**Supplementary Note** online). In all seven case-control groups, rs7501939 C and rs4430796 A showed a protective effect against the disease (that is, an OR < 1.0). Combining results from all eight T2D case-control groups, including the Icelandic group, gave an OR of 0.91 ($P = 9.2 \times 10^{-7}$) for rs7501939 C and an OR of 0.91 ($P = 2.7 \times 10^{-7}$) for rs4430796 A (**Table 5**). In a joint analysis, the effect of rs4430796 remained significant with adjustment for rs7501939 ($P = 0.016$), whereas rs7501939 did not after adjusting for rs4430796 ($P = 0.41$). We note that the former was mainly driven by the data from West Africa, where the correlation between the two

Table 3 Association results for SNPs on 17q24.3 and prostate cancer in Iceland, The Netherlands, Spain and the US

	Frequency			
Study population (*N* cases/*N* controls) and variant (allele)	Cases	Controls	OR (95% c.i.)	*P* value
Iceland (1,501/11,290)				
rs1859962 (G)	0.489	0.453	1.16 (1.07–1.26)	3.1×10^{-4}
rs7214479 (T)	0.451	0.415	1.16 (1.07–1.26)	3.3×10^{-4}
rs6501455 (A)	0.538	0.501	1.16 (1.07–1.26)	3.0×10^{-4}
rs983085 (C)[a]	0.542	0.504	1.16 (1.07–1.26)	2.0×10^{-4}
The Netherlands (999/1,466)				
rs1859962 (G)	0.522	0.456	1.30 (1.16–1.46)	6.8×10^{-6}
rs7214479 (T)	0.474	0.428	1.20 (1.07–1.35)	1.5×10^{-3}
rs6501455 (A)	0.544	0.488	1.25 (1.12–1.40)	1.1×10^{-4}
Spain (456/1,078)				
rs1859962 (G)	0.512	0.476	1.15 (0.99–1.35)	0.071
rs7214479 (T)	0.455	0.426	1.13 (0.96–1.32)	0.14
rs6501455 (A)	0.581	0.552	1.13 (0.97–1.32)	0.13
Chicago (537/510)				
rs1859962 (G)	0.513	0.456	1.25 (1.06–1.49)	9.8×10^{-3}
rs7214479 (T)	0.460	0.416	1.20 (1.01–1.42)	0.041
rs6501455 (A)	0.549	0.586	0.86 (0.72–1.02)	0.083
All excluding Iceland (1,992/3,054)[b]				
rs1859962 (G)	–	0.463	1.25 (1.15–1.35)	8.3×10^{-8}
rs7214479 (T)	–	0.423	1.18 (1.09–1.28)	7.0×10^{-5}
rs6501455 (A)	–	0.542	1.12 (1.05–1.20)	6.2×10^{-3}
All combined (3,493/14,344)[b]				
rs1859962 (G)	–	0.460	1.20 (1.14–1.27)	2.5×10^{-10}
rs7214479 (T)	–	0.421	1.17 (1.10–1.24)	8.1×10^{-8}
rs6501455 (A)	–	0.532	1.14 (1.08–1.21)	6.9×10^{-6}

All *P* values shown are two sided. Shown are the numbers of cases and controls (*N*), allelic frequencies of variants in affected and control individuals, the allelic odds ratio (OR) with 95% confidence interval (95% c.i.) and *P* values based on the multiplicative model.
[a]SNPs rs983085 and rs6501455 were almost perfectly correlated (r^2 = 0.99), but rs983085 failed in genotyping in the non-Icelandic groups. [b]For the combined study populations, the reported control frequency was the average, unweighted control frequency of the individual populations, whereas the OR and the *P* values were estimated using the Mantel-Haenszel model.

SNPs is substantially lower than in individuals of European ancestry (r^2 = 0.22 and r^2 = 0.77 in the Yoruba and CEU HapMap samples, respectively). For T2D, a recent report[15] describes similar findings (OR = 0.89, $P = 5 \times 10^{-6}$) for allele G of the SNP rs757210, which is substantially correlated with rs4430796 A (D′ = 0.96; r^2 = 0.62; based on the CEU HapMap data set). This reinforces the finding that one or more variants in *TCF2* that confer risk of prostate cancer are protective against T2D. Notably, removing individuals with T2D from the Icelandic case-control group had minimal impact on the association of rs4430796 with prostate cancer (**Supplementary Note**).

The more distal SNP, rs1859962, on chromosome 17q24.3 is in a 177.5-kb LD block spanning positions 66.579 Mb to 66.757 Mb

Table 4 Model-free estimates of the genotype OR of rs4430796 (A) at 17q12 and rs1859962 (G) at 17q24.3

		Genotype OR[a]					
Study group and variant (allele)	Allelic OR	00	0X (95% c.i.)	XX (95% c.i.)	*P* value[b]	*P* value[c]	PAR
Iceland							
rs4430796 (A)	1.20	1	1.12 (0.97–1.29)	1.40 (1.19–1.64)	0.31	8.3×10^{-5}	0.14
rs1859962 (G)	1.16	1	1.35 (1.18–1.54)	1.33 (1.13–1.57)	3.4×10^{-3}	2.3×10^{-5}	0.19
All except Iceland							
rs4430796 (A)	1.24	1	1.34 (1.18–1.52)	1.56 (1.32–1.84)	0.12	4.5×10^{-7}	0.23
rs1859962 (G)	1.25	1	1.32 (1.17–1.49)	1.57 (1.33–1.84)	0.24	2.9×10^{-7}	0.22
All combined							
rs4430796 (A)	1.22	1	1.24 (1.13–1.36)	1.48 (1.32–1.66)	0.57	2.0×10^{-10}	0.19
rs1859962 (G)	1.20	1	1.33 (1.21–1.44)	1.45 (1.29–1.62)	6.0×10^{-3}	5.1×10^{-11}	0.21

PAR, population attributable risk; OR, odds ratio; 95% c.i., 95% confidence interval.
[a]Genotype odds ratios for heterozygous (0X) and homozygous carriers (XX) compared with non-carriers (00). [b]Test of the multiplicative model (the null hypothesis) versus the full model (one degree of freedom). [c]Test of no effect (the null hypothesis) versus the full model (two degrees of freedom).

Table 5 Association results for SNPs in the *TCF2* gene on 17q12 and type 2 diabetes

	Frequency			
Study population (*N* cases/*N* controls) and variant (allele)	Cases	Controls	OR (95% c.i.)	*P* value
Iceland[a] (1,380/9,940)				
rs7501939 (C)	0.549	0.582	0.88 (0.80–0.96)	0.0045
rs4430796 (A)	0.482	0.521	0.86 (0.78–0.95)	0.0021
Denmark A (264/596)				
rs7501939 (C)	0.525	0.593	0.76 (0.62–0.93)	0.0088
rs4430796 (A)	0.452	0.530	0.73 (0.60–0.90)	0.0032
Denmark B (1,365/4,843)				
rs7501939 (C)	0.579	0.596	0.93 (0.85–1.02)	0.11
rs4430796 (A)	0.507	0.528	0.92 (0.85–1.00)	0.062
Philadelphia (457/967)				
rs7501939 (C)	0.569	0.613	0.83 (0.71–0.98)	0.028
rs4430796 (A)	0.477	0.527	0.82 (0.70–0.96)	0.013
Scotland (3,741/3,718)				
rs7501939 (C)	0.607	0.615	0.97 (0.91–1.03)	0.31
rs4430796 (A)	0.517	0.526	0.97 (0.91–1.03)	0.29
The Netherlands (367/915)				
rs7501939 (C)	0.563	0.579	0.94 (0.79–1.11)	0.46
rs4430796 (A)	0.494	0.506	0.95 (0.79–1.14)	0.58
Hong Kong (1,495/993)				
rs7501939 (C)	0.768	0.791	0.87 (0.76–1.00)	0.054
rs4430796 (A)	0.731	0.754	0.89 (0.78–1.01)	0.073
West Africa[b] (867/1,115)				
rs7501939 (C)	0.400	0.437	0.87 (0.77–0.99)	0.042
rs4430796 (A)	0.271	0.313	0.80 (0.69–0.92)	0.0024
All groups excluding Iceland				
rs7501939 (C)	–	–	0.91 (0.87–0.95)	3.4×10^{-5}
rs4430796 (A)	–	–	0.92 (0.88–0.95)	1.8×10^{-5}
All groups combined (9,936/23,087)				
rs7501939 (C)	–	–	0.91 (0.87–0.94)	9.2×10^{-7}
rs4430796 (A)	–	–	0.91 (0.87–0.94)	2.7×10^{-7}

All *P* values shown are two sided. Shown are the numbers of cases and controls (*N*), allelic frequencies of variants in affected and control individuals, the allelic odds ratio (OR) with 95% confidence interval (95% c.i.) and *P* values based on the multiplicative model.
[a]Men known to have prostate cancer were excluded from the Icelandic T2D group (both affected individuals and controls). [b]Results for the five West African tribes have been combined using a Mantel-Haenszel method. The frequency of the variant in West African affected individuals and controls is the weighted average over the five tribes.

(National Center for Biotechnology Information (NCBI) build 35), based on the CEU HapMap group. The closest telomeric gene is *SOX9*, located ~900 kb away from the LD block. One mRNA (BC039327) and several unspliced ESTs have been localized to this region, but it does not contain any known genes (University of California Santa Cruz Genome Browser, May 2004 assembly). RT-PCR analysis of various cDNA libraries, including those derived from the prostate, detected expression of the BC039327 mRNA only in a testis library (data not shown), in line with previously reported results[16].

In summary, we have found that two common variants on chromosome 17q, rs4430796 A and rs1859962 G, contribute to the risk of prostate cancer in four populations of European descent. Together, based on the combined results, these two variants have an estimated joint population attributable risk (PAR) of ~36%, which is substantial from a public health viewpoint. The large PAR is a consequence of the high frequencies of these variants. However, as their relative risks, as estimated by the ORs, are not high, the sibling risk ratio[17] that they account for is only ~1.009 for each variant separately and ~1.018 jointly. As a consequence, they can explain only a small fraction of the familial clustering of the disease and can therefore generate only modest linkage scores. We were most intrigued that the variant in *TCF2* is associated with increased risk of prostate cancer but reduced risk of T2D in individuals of European, African and Asian descent. The discovery of a sequence variant in the *TCF2* gene that accounts for at least part of the inverse relationship between these two diseases provides a step toward understanding the complex biochemical checks and balances that result from the pleiotropic impact of singular genetic variants. Previous explanations of the well-established inverse relationship between prostate cancer and T2D have centered on the impact of the metabolic and hormonal environment of diabetic men. However, we note that the protective effect of both the *TCF2* SNPs against T2D is too modest for its impact on prostate cancer risk to be merely a by-product of its impact on T2D. Indeed, we favor the notion that the primary functional impact of rs4430796 (or a presently unknown correlated variant) is on one or more metabolic or hormonal pathways important for the normal functioning of individuals throughout their lives that incidentally modulate the risk of developing prostate cancer and T2D late in life.

METHODS

Icelandic study population. Men diagnosed with prostate cancer were identified based on a nationwide list from the Icelandic Cancer Registry (ICR) that contained all 3,886 Icelandic prostate cancer patients diagnosed from January 1, 1955, to December 31, 2005. The Icelandic prostate cancer sample collection included 1,615 patients (diagnosed from December 1974 to December 2005) who were recruited from November 2000 until June 2006 out of the 1,968 affected individuals who were alive during the study period (a participation rate of about 82%). A total of 1,541 affected individuals were included in a genome-wide SNP genotyping effort, using the Infinium II assay method and the Illumina Sentrix HumanHap300 BeadChip. Of these, 1,501 (97%) were successfully genotyped according to our quality control criteria (**Supplementary Methods** online) and were used in the present case-control association analysis. The mean age at diagnosis for the consenting patients was 71 years (median 71 years; range, 40–96 years), and the mean age at diagnosis was 73 years for all individuals with prostate cancer in the ICR. The median time from diagnosis to blood sampling was 2 years (range, 0–26 years) (see ref. 1 for a more detailed description of the Icelandic prostate cancer study population). No significant difference was seen in frequencies of rs7501939 (C), rs4430796 (A) or rs1859962 (G) between men diagnosed before 1998 and those diagnosed in 1998 or later ($P = 0.74$, $P = 0.87$ and $P = 0.35$, respectively). More specifically, using only cases diagnosed in 1998 or later ($N = 880$) versus all our controls ($N = 11,289$), we obtained OR values of 1.16 ($P = 0.004$), 1.20 (5.5×10^{-4}) and 1.20 (5×10^{-4}) for rs7501939 (C), rs4430796 (A) and rs1859962 (G), respectively. The 11,290 controls (5,010 males and 6,280 females) used in this study consisted of 758 controls randomly selected from the Icelandic genealogical database and 10,532 individuals from other ongoing genome-wide association studies at deCODE (specifically, ~1,400 from studies on T2D, ~1,600 from studies on breast cancer and 1,800 from studies on myocardial infarction; studies on colon cancer, anxiety, addiction, schizophrenia and infectious diseases provided ~700–1,000 controls each). The controls had a mean age of 66 years (median, 67 years; range, 22–102 years). The male controls were absent from the ICR's nationwide list of prostate cancer patients.

The study was approved by the Data Protection Commission of Iceland and the National Bioethics Committee of Iceland. Written informed consent was obtained from all patients, relatives and controls. Personal identifiers associated with medical information and blood samples were encrypted with a third-party encryption system as previously described[18].

Study populations from The Netherlands, Spain and the US. The total number of men with prostate cancer from the Netherlands in this study was 1,013, of whom 999 (98%) were successfully genotyped. This study population comprised two recruitment sets of men with prostate cancer: Group A, comprising 390 hospital-based affected individuals recruited from January 1999 to June 2006 at the Urology Outpatient Clinic of the Radboud University Nijmegen Medical Centre (RUNMC), and Group B, consisting of 623 affected individuals recruited from June 2006 to December 2006 through a population-based cancer registry held by the Comprehensive Cancer Centre East. Both groups were of self-reported European descent. The average age at diagnosis for patients in Group A was 63 years (median, 63 years; range, 43–83 years). The average age at diagnosis for patients in Group B was 65 years (median 66 years; range, 43–75 years).

The 1,466 control individuals from The Netherlands were cancer free and were matched for age with the cases. They were recruited as part of the Nijmegen Biomedical Study, a population-based survey conducted by the Department of Epidemiology and Biostatistics and the Department of Clinical Chemistry of the RUNMC, in which 9,371 individuals participated from a total of 22,500 age- and sex-stratified randomly selected inhabitants of Nijmegen, The Netherlands. Control individuals from the Nijmegen Biomedical Study were invited to participate in a study on gene-environment interactions in multifactorial diseases such as cancer. All the 1,466 participants in the present study are of self-reported European descent and were fully informed about the goals and the procedures of the study. The study protocol was approved by the Institutional Review Board of Radboud University, and all study subjects gave written informed consent.

The Spanish study population consisted of 464 men with prostate cancer, of whom 456 (98%) were successfully genotyped. The cases were recruited from the Oncology Department of Zaragoza Hospital in Zaragoza, Spain, from June 2005 to September 2006. All were of self-reported European descent. Clinical information, including age at onset, grade and stage, was obtained from medical records. The average age at diagnosis for the patients was 69 years (median, 70 years; range, 44–83 years). The 1,078 Spanish control individuals were approached at Zaragoza University Hospital and were confirmed to be prostate cancer free before they were included in the study. Study protocols were approved by the Institutional Review Board of Zaragoza University Hospital. All subjects gave written informed consent.

The Chicago study population consisted of 557 men with prostate cancer, of whom 537 (96%) were successfully genotyped. The affected individuals were recruited from the Pathology Core of Northwestern University's Prostate Cancer Specialized Program of Research Excellence (SPORE) from May 2002 to September 2006. The average age at diagnosis for the affected individuals was 60 years (median, 59 years; range, 39–87 years). The 514 European American controls were recruited as healthy control subjects for genetic studies at the University of Chicago and Northwestern University Medical School. Study protocols were approved by the Institutional Review Boards of Northwestern University and the University of Chicago. All subjects gave written informed consent.

For description of the diabetes case-control groups, see the **Supplementary Note.**

Association analysis. All Icelandic case and control samples were assayed with the Illumina Infinium HumanHap300 SNP chip. This chip contains 317,503 SNPs and provides about 75% genomic coverage in the Utah CEPH (CEU) HapMap samples for common SNPs at $r^2 \geq 0.8$. For the association analysis, 310,520 SNPs were used; 6,983 SNPs were deemed unusable owing to reasons such as monomorphism, low yield (<95%) and failure of Hardy-Weinberg equilibrium (HWE) (**Supplementary Methods**). Samples with a call rate <98% were excluded from the analysis. Single-SNP genotyping for the five SNPs reported here and the four case-control groups was carried out by deCODE Genetics, applying the Centaurus[19] (Nanogen) platform to all populations studied (**Supplementary Methods** and **Supplementary Table 2a** online). For the five SNPs genotyped by both methods in 1,501 affected individuals and 758 controls from Iceland, the concordance rate for genotypes was >99.5% between the Illumina platform and the Centaurus platform.

For SNPs that were in strong LD, whenever the genotype of one SNP was missing for an individual, the genotype of the correlated SNP was used to provide partial information through a likelihood approach, as we have done before[1]. This ensured that results presented in **Tables 2–5** were always based on the same number of individuals, allowing meaningful comparisons of results for highly correlated SNPs. A likelihood procedure described in a previous publication[20] and implemented in NEMO software was used for the association analyses. We attempted to genotype all individuals and all SNPs reported in **Tables 2–5**. For each SNP, the yield was >95% in every group. The only exception was in the case of refinement marker rs4430796, which was not a part of the HumanHap 300 chip. For this SNP, using a single SNP assay to genotype, we attempted to genotype 1,883 of the 11,290 Icelandic controls (genotyping was successful for 99% of them (1,860 individuals)) as well as all affected Icelandic individuals and all individuals from the replication study groups. Most notably, for the 17q12 locus, when we evaluated the significance of one SNP (for example, rs4430796, rs7501939 or rs3760511) with adjustment for one or two other SNPs, whether we used all 11,289 Icelandic controls that had genotypes for at least one of the three markers in **Table 2** and handled the missing data by applying a likelihood approach as mentioned above or whether we applied logistic regression only to individuals that had genotypes for all three markers, the resulting P values are very similar. We tested the association of an allele with prostate cancer using a standard likelihood ratio statistic that, if the subjects were unrelated, would have asymptotically a χ^2 distribution with one degree of freedom under the null hypothesis. Allelic frequencies rather than carrier frequencies are presented for the markers in the main text, but genotype counts are provided in **Supplementary Table 3** online. Allele-specific ORs and associated P values were calculated assuming a multiplicative model for the two chromosomes of an individual[21]. For each of the four case-control groups, there was no significant deviation from HWE in the controls ($P > 0.01$). When estimating genotype-specific OR (**Table 3**), we estimated genotype frequencies in the population assuming HWE. We feel that this estimate is more stable than an estimate calculated using the observed genotype counts in controls directly. However, we note that these two approaches gave very similar estimates in this instance. Results from multiple case-control groups were combined using a Mantel-Haenszel model[22] in which the groups were allowed to have different population frequencies for alleles, haplotypes and genotypes but were assumed to have common relative risks. All four of the European sample groups include both male and female controls. We did not detect a significant difference between male and female controls for SNPs in **Tables 2–4** for each of the groups after correction for the number of tests performed. We note that for all the three significant variants (rs7501939, rs4430796 and rs1859962) reported in **Tables 2** and **3**, we did not detect any significant differences in frequencies among the different groups of affected individuals (see description of Icelandic control samples) that make up the Icelandic genome-wide control sets ($P = 0.30$, 0.55 and 0.88, respectively). The individuals with T2D were removed when this test was performed for rs7501939 and rs4430796. Our analysis of the data does not indicate any differential association by gender of rs7501939 or rs4430796 to T2D. We used linear regression to estimate the relationship between age at onset for prostate cancer and number of copies of at-risk alleles (for rs7501939 and rs1859962) carried by affected individuals, using group as an indicator.

To investigate potential interaction between rs7501939 C and rs1859962 G located at 17q12 and 17q24.3, respectively, we performed two analyses. First, we checked for the absence of significant correlation between those alleles among cases. Second, using logistic regression, we demonstrated that the interaction term was not significant ($P = 0.57$). The joint PAR was calculated as $1 - ((1 - \mathrm{PAR}_1) \times (1 - \mathrm{PAR}_2))$, where PAR_1 and PAR_2 are the individual PARs for each SNP calculated under the full model and assuming no interaction between the SNPs.

We note that for the SNP rs757210, others have reported the results for allele A[15]. However, in the main text, we provide their corresponding results for the other allele (allele G of rs757210) because that allele was the one positively correlated with our reported allele C of rs7501939.

Correction for relatedness and genomic control. Some individuals in the Icelandic case-control groups were related to each other, causing the aforementioned χ^2 test statistic to have a mean >1. We estimated the inflation factor by calculating the mean of the 310,520 χ^2 statistics, which is 1.098. Using a method of genomic control[23] to adjust for both relatedness and potential population stratification, results presented here are based on adjusting the χ^2 statistics by dividing each of them by 1.098. **Supplementary Figure 1** is a Q-Q plot of the observed χ^2 statistics, before and after adjustment, against the χ^2 distribution with one degree of freedom.

URLs. Cancer Genetic Markers of Susceptibility Project: http://cgems.cancer.gov/. University of California Santa Cruz Genome Browser: http://www.genome.ucsc.edu.

Requests for materials: kstefans@decode.is or julius.gudmundsson@decode.is

ACKNOWLEDGMENTS

We thank the research subjects whose contribution made this work possible. We also thank the nurses at Noatun (deCODE's sample recruitment center) and personnel at the deCODE core facilities. This project was funded in part by contract number 018827 (Polygene) from the 6th Framework Program of the European Union and by US Department of Defense Congressionally Directed Medical Research Program W81XWH-05-1-0074. Support for the Africa America Diabetes Mellitus (AADM) study is provided by multiple institutes of the US National Institutes of Health: the National Center on Minority Health and Health Disparities (3T37TW00041-03S2), National Institute of Diabetes and Digestive and Kidney (DK072128), the National Human Genome Research Institute and the National Center for Research Resources (RR03048). The Scottish Diabetes case control study was funded by the Wellcome Trust. C.N.A.P. and A.M. are supported by the Scottish Executive Generation Scotland Initiative. The Hong Kong Diabetes case control study was supported by the Hong Kong Research Grants Committee Central Allocation Scheme CUHK 1/04C.

AUTHOR CONTRIBUTIONS

The principal investigators of the prostate cancer replication study samples are J.I.M., L.A.K. and W.J.C.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Amundadottir, L.T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**, 652–658 (2006).
2. Gudmundsson, J. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).
3. Haiman, C.A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39**, 638–644 (2007).
4. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
5. Roeder, K., Bacanu, S.A., Wasserman, L. & Devlin, B. Using linkage genome scans to improve power of association in genome scans. *Am. J. Hum. Genet.* **78**, 243–252 (2006).
6. Lange, E.M. *et al.* Genome-wide scan for prostate cancer susceptibility genes using families from the University of Michigan prostate cancer genetics project finds evidence for linkage on chromosome 17 near BRCA1. *Prostate* **57**, 326–334 (2003).
7. Xu, J. *et al.* A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics. *Am. J. Hum. Genet.* **77**, 219–229 (2005).
8. Lange, E.M. *et al.* Fine-mapping the putative chromosome 17q21–22 prostate cancer susceptibility gene to a 10 cM region based on linkage analysis. *Hum. Genet.* **121**, 49–55 (2007).
9. Zuhlke, K.A. *et al.* Truncating BRCA1 mutations are uncommon in a cohort of hereditary prostate cancer families with evidence of linkage to 17q markers. *Clin. Cancer Res.* **10**, 5975–5980 (2004).
10. Kraft, P. *et al.* Genetic variation in the HSD17B1 gene and risk of prostate cancer. *PLoS Genet* **1**, e68 (2005).
11. White, K.A., Lange, E.M., Ray, A.M., Wojno, K.J. & Cooney, K.A. Prohibitin mutations are uncommon in prostate cancer families linked to chromosome 17q. *Prostate Cancer Prostatic Dis.* **9**, 298–302 (2006).
12. Bellanne-Chantelot, C. *et al.* Large genomic rearrangements in the hepatocyte nuclear factor-1beta (TCF2) gene are the most frequent cause of maturity-onset diabetes of the young type 5. *Diabetes* **54**, 3126–3132 (2005).
13. Edghill, E.L., Bingham, C., Ellard, S. & Hattersley, A.T. Mutations in hepatocyte nuclear factor-1beta and their related phenotypes. *J. Med. Genet.* **43**, 84–90 (2006).
14. Kasper, J.S. & Giovannucci, E. A meta-analysis of diabetes mellitus and the risk of prostate cancer. *Cancer Epidemiol. Biomarkers Prev.* **15**, 2056–2062 (2006).
15. Winckler, W. *et al.* Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes. *Diabetes* **56**, 685–693 (2007).
16. HillHarfe, K.L. *et al.* Fine mapping of chromosome 17 translocation breakpoints > or = 900 Kb upstream of SOX9 in acampomelic campomelic dysplasia and a mild, familial skeletal dysplasia. *Am. J. Hum. Genet.* **76**, 663–671 (2005).
17. Risch, N. Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.* **46**, 222–228 (1990).
18. Gulcher, J.R., Kristjansson, K., Gudbjartsson, H. & Stefansson, K. Protection of privacy by third-party encryption in genetic research in Iceland. *Eur. J. Hum. Genet.* **8**, 739–742 (2000).
19. Kutyavin, I.V. *et al.* A novel endonuclease IV post-PCR genotyping system. *Nucleic Acids Res.* **34**, e128 (2006).
20. Gretarsdottir, S. *et al.* The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat. Genet.* **35**, 131–138 (2003).
21. Falk, C.T. & Rubinstein, P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* **51**, 227–233 (1987).
22. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22**, 719–748 (1959).
23. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).

Ann. Hum. Genet. (1987), 51, 227–233
Printed in Great Britain

Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations

C. T. FALK AND P. RUBINSTEIN

The Lindsley F. Kimball Research Institute of The New York Blood Center, 310 E. 67th St., New York, NY 10021

SUMMARY

An alternative to Woolf's (1955) relative risk (RR) statistic is proposed for use in calculating the risk of disease in the presence of particular antigens or phenotypes. This alternative uses, as the control sample, the parental antigens or haplotypes not present in the affected child. The formulation of a haplotype relative risk (HRR) thus eliminates the problems of sampling from the same homogeneous population to form both the disease sample and an appropriate control.

We show that, in families selected through a single affected individual, where transmission of the four parental haplotypes can be followed unambiguously, the mathematical expectation of the HRR is identical to that of the RR. Since the sample formed from the 'non-affected' parental haplotypes is clearly from the same population as the disease sample, the HRR thus provides a reliable alternative to the RR. A further advantage obtains when family data are being collected as part of a study since the control sample is then automatically contained in the family material.

Data from studies of patients with insulin dependent diabetes mellitus (IDDM) are used to obtain an estimate of the risk to those with HLA antigens or phenotypes associated with IDDM using the HRR statistic. A comparison of the HRR's and RR's for these data is also presented.

INTRODUCTION

Relative risks have been used for some time to estimate the increased risk of contracting a disease, given that a certain condition (or trait) is present, over that of the group lacking the condition. This formal definition of a relative risk requires prospective information that is not easily obtained and the relative risk is often approximated by the more easily obtained cross product

$$\frac{\Pr(Q\,|\,\text{aff})\Pr(q\,|\,\text{control})}{\Pr(q\,|\,\text{aff})\Pr(Q\,|\,\text{control})},$$

where Q stands for the presence of the condition or trait and q for the lack of the condition, and the four terms are conditional probabilities as indicated. When the overall frequency of the disease in a population is low, this estimate will closely approximate the true relative risk. This odds ratio was proposed by Woolf (1955) to estimate the risk of contracting either peptic ulcers or stomach cancer for individuals of particular ABO phenotypes. Since then it has been used to calculate risks for genetic markers associated with many diseases and its most notable use has been in studying several HLA-associated diseases such as insulin dependent diabetes mellitus (IDDM), coeliac disease, multiple sclerosis and ankylosing spondylitis. Several assumptions are generally made about the underlying population from which both the disease sample and the control sample are obtained, most importantly that both samples are drawn from the same genetically homogeneous population in an unbiased way. By this we mean that the disease sample should be selected on a clear-cut ascertainment criterion, e.g. randomly chosen affected individuals with no bias pertaining to other factors, and the control sample should be a strictly random sample from the same genetic population. In practice, this latter criterion is rather difficult to fulfil and most often the control is created from conveniently available data drawn from a population thought to be somewhat closely related to that from which the disease sample was drawn.

Several years ago we proposed (Rubinstein *et al.* 1981) an alternative method for obtaining the control sample for relative risk (RR) estimations that eliminated the problems of sampling from a single homogeneous population. This method used, as a control, those parental haplotypes not present in the affected child and was therefore called the haplotype relative risk (HRR). This method has several appealing features including freedom from collection of proper control samples. Additionally, where families are to be studied anyway, collection of the family data automatically includes collection of the necessary control sample. It is, however, necessary to demonstrate that the HRR estimate has the appropriate characteristics. In this paper we will show that, assuming the 'ideal' conditions inherent in the definition of RR, namely, control and disease samples both randomly chosen from the same homogeneous random mating population, the expected value of the HRR is identical to that of the conventional RR. We will then illustrate its use in the estimation of risks for HLA antigens and phenotypes associated with IDDM.

THE MODEL

Consider a set of families that has been ascertained through a single affected child, where the relevant disease locus is closely linked to a normal polymorphic genetic marker (e.g. HLA) and where certain alleles (antigens) are associated with the disease. For purposes of concreteness, we will assume that the disease is recessively inherited, although the same arguments hold for dominance and for other inheritance models as well. Assume that the HLA haplotypes present in the parents can be followed unambiguously in transmission to the offspring and designate the two inherited by the affected child as 'a' (paternal) and 'c' (maternal). Thus haplotypes a and c are assumed to carry the disease allele, say 'n'. In the special case where the child as well as both parents are ac, it is not certain whether the child gets the a from the mother or the father. However, it is still known that one a and one c haplotype were transmitted to the affected child, and thus carry the n allele, and that the haplotypes not passed on to the affected child were also a and c. The latter can therefore be included in the 'random sample' as described below. Now if we have truly obtained our sample as a random, singly selected sample, the two parental haplotypes not transmitted to the affected child (say b and d) will represent a random sample of haplotypes from the population at large and will thus carry the disease allele (n) or the normal allele (N) with probabilities equal to the allele frequencies in the population (say p_1 and p_2, respectively, $p_1+p_2 = 1$). The validity of this observation requires compliance with certain other assumptions including (1) that the parents are not inbred, (2) that there is no correlation within or between parental phenotypes and (3) that there is no differential fertility of the disease phenotypes.

Now assume that an antigen 'Q' at the HLA locus is in positive linkage disequilibrium with n, the disease allele. We wish to calculate the relative risk to carriers of Q of contracting the disease. We will use as our control population the set of '*b*' and '*d*' haplotypes from our sample of disease families (that is, those haplotypes within a family not carried by the single affected proband). Using this control we will then calculate the conventional cross product odds ratio given above to obtain the haplotype relative risk (HRR). Define the relevant population frequencies as follows:

$$
\begin{aligned}
f(Q) &= q_1,\\
f(q) &= q_2 = 1-q_1 \quad \text{(where } q \text{ represents all other alleles)},\\
f(n) &= p_1,\\
f(N) &= p_2 = 1-p_1,\\
f(Qn) &= x_1 = p_1 q_1+\delta,\\
f(QN) &= x_2 = p_2 q_1-\delta,\\
f(qn) &= x_3 = p_1 q_2-\delta,\\
f(qN) &= x_4 = p_2 q_2+\delta,
\end{aligned}
$$

where δ is the measure of disequilibrium between n and Q.

We now need the four conditional probabilities necessary for the odds ratio. For the affected sample these are the same, regardless of how we choose our control.

$$
\begin{aligned}
\Pr(Q\,|\,\text{aff}) &= \frac{x_1^2+2x_1x_3}{(x_1+x_3)^2},\\
&= \frac{p_1^2-x_3^2}{p_1^2},
\end{aligned}
$$

$$
\Pr(\text{not}\,Q\,|\,\text{aff}) = \frac{x_3^2}{(x_1+x_3)^2} = \frac{x_3^2}{p_1^2}.
$$

Now since the control haplotypes will be a random sample from the population, the conditional probabilities will be:

$$
\Pr(Q\,|\,\text{control}) = 1-(x_3+x_4)^2 = 1-q_2^2
$$

$$
\Pr(\text{not}\,Q\,|\,\text{control}) = (x_3+x_4)^2 = q_2^2.
$$

Thus the estimate of the HRR is:

$$
\begin{aligned}
\text{HRR} &= \frac{\Pr(Q\,|\,\text{aff})\Pr(\text{not}\,Q\,|\text{control})}{\Pr(\text{not}\,Q\,|\,\text{aff})\Pr(Q\,|\,\text{control})}\\
&= \frac{(p_1^2-x_3^2)\,q_2^2}{x_3^2(1-q_2^2)},
\end{aligned}
$$

which is identical to the equivalent expression for the conventional RR.

EXAMPLE

Using data collected for the 9th HLA Workshop (Bertrams & Baur, 1984) we looked at the sample of families, submitted for study, where a single child was affected with IDDM and where the ethnic background was caucasoid (Western European or North American). The patients

Table 1. DR *phenotypes of* IDDM *disease sample, simplex cases*

DR type	No. obs.	No. exp.
DR3, 3	6	7·8
DR3, 4	25	18·2
DR4, 4	4	10·7
DR3, X	16	19·1
DR4, X	29	22·4
DRX, X	10	11·7
Total	90	89·9

$p(DR3) = 0{\cdot}294$; $p(DR4) = 0{\cdot}344$; $p(DRX) = 0{\cdot}361$; $\alpha/\beta = (0{\cdot}278)/(0{\cdot}202) = 1{\cdot}38$.

Table 2. DR *phenotypes of 'control' sample consisting of non-affected parental haplotypes*

DR type	No. obs.	No. exp.
DR3, 3	0	0·77
DR3, 4	2	1·23
DR4, 4	0	0·49
DR3, X	13	12·26
DR4, X	10	9·76
DRX, X	48	48·49
Total	73	73·00

$p(DR3) = 0{\cdot}103$; $p(DR4) = 0{\cdot}082$; $P(DRX) = 0{\cdot}815$; $\chi^2 = 1{\cdot}79$, 2. d.f.

were categorized with respect to their HLA DR phenotypes using three distinct allelic groups *DR3*, *DR4*, and *DRX*, where *DRX* represents all other *DR* antigens except *DR3* and *DR4*. The results are shown in Table 1 with estimated allele frequencies and observed and 'Hardy–Weinberg expected' numbers for each phenotypic class. The α/β ratio of Falk *et al.* (1983) was also calculated and found to be 1·38. This ratio relates the observed frequency (α) of, say the DR3.4 phenotype, to the Hardy–Weinberg expected frequency [$\beta = 2p(DR3)p(DR4)$] in a sample of diseased individuals (Table 1). A value in excess of 1·0 is an indication that the associated susceptibility locus does not show a simple dominant or recessive mode of inheritance with a single susceptibility allele. The value of 1·38 found here is characteristic of samples of IDDM individuals where an excess of DR3, 4's is often observed thus suggesting a more complex mode of inheritance for susceptibility (Falk, 1984). The 'control group' was made up of the parental haplotype pairs not present in the affected child (only families in which all four HLA haplotypes could be followed were used). There were 146 parental control haplotypes. The allele frequencies for *DR3*, *DR4*, and *DRX* in this group were 0·103, 0·082, and 0·815 respectively. These values agree remarkably well with the total frequencies obtained for the 'random mating population' comprising all caucasoid random individuals submitted to the 9th HLA Workshop (Baur *et al.* 1984) (see, e.g. the table on page 694, where the *DR* marginal frequencies are 0·122, 0·129, and 0·749 for the same three *DR* alleles). If the control haplotypes from each family are assumed to be a 'control individual', we obtain a control population sample of 73 which is in H–W equilibrium ($\chi^2 = 1{\cdot}79$, 2 d.f., see Table 2).

In Table 3, we compare the HRR's for DR3 and DR4 to the RR's calculated using a 'contrived control population' from the 9th HLA Workshop population data referred to above. This 'population' is assumed to be in H–W equilibrium and our 'random sample' is of the same

Table 3. *HRR's and RR's for the DR3 and DR4 antigens in a sample of simplex IDDM patients*

(The control for the HRR's is the sample of parental haplotypes not present in the affected individuals. The control for the RR's was obtained by 'creating' a H-W sample assuming the antigen frequencies recorded for the 9th HLA workshop (Baur *et al.* 1984).)

	HRR				RR		
	DR3				*DR3*		
	+	−			+	−	
Disease	47	43	90	Disease	47	43	90
control	15	58	73	control	21	69	90
	62	101	163		68	112	180
	HRR = 4·23				RR = 3·59		
	$p = 2{\cdot}6 \times 10^{-5}$				$p = 5{\cdot}3 \times 10^{-5}$		
	DR4				*DR4*		
	+	−			+	−	
Disease	58	32	90	Disease	58	32	90
control	12	61	73	control	22	68	90
	70	93	163		80	100	180
	HRR = 9·21				RR = 5·60		
	$p = 7{\cdot}6 \times 10^{-10}$				$p = 6{\cdot}8 \times 10^{-8}$		

Table 4. *HRR's and RR's for the DR3, 3, DR3, 4 and DR4, 4 phenotypes*

(Samples are the same as those described in Table 3. In each case comparison is made relative to the 'base group' DRX, X to avoid the problems of non-independent risk estimates.).

DR type	Disease sample	Parental control	Workshop control
DR3, 3	6	0	1·3
DR3, 4	25	2	2·8
DR4, 4	4	0	1·5
DR3, X	16	13	16·4
DR4, X	29	10	17·4
DRX, X	10	48	50·5
Total	90	73	89·9

HRR	RR
HRR(3, 4) = 60·0	RR(3, 4) = 45·1
HRR(3, 3) = undefined	RR(3, 3) = 23·3
HRR(4, 4) = undefined	RR(4, 4) = 13·5

If 'expected values' are substituted for the zero observations in the parental control, one gets:

HRR′(3, 3) = 37·4,
HRR′(4, 4) = 39·2.

size as our disease sample (i.e. 90 individuals). Table 4 gives HRR's and RR's for the three DR phenotypes DR3, 3, DR3, 4, and DR4, 4 using the same samples. Here the risks are compared to the baseline phenotype DRX, X in each case since the risks are not independent (cf. Curie-Cohen, 1981, Svejgaard & Ryder, 1981). Note that the HRR's for DR3, 3 and DR4, 4 are undefined since there are no 'individuals' with those phenotypes in the control sample of 73. If expected values are substituted for the 'zero' values in those cases HRR's can be estimated as given at the bottom of Table 4, but the use of such estimates must be made with caution.

DISCUSSION

One of the major problems inherent in proper calculations of relative risks (RR's) is that of choosing an appropriate control. A basic assumption in the use of RR's is that both the affected sample and the control sample are chosen at random from the same genetically homogeneous random mating population with no selection criteria except for the disease status required for inclusion in the affected sample. In practice this is a difficult criterion to fulfil. Additionally, it adds a significant amount of work to select and test such a control sample. It is therefore often assumed that the control sample is simply a hypothetical sample created from a population thought to be similar to that of the disease sample and 'generated' from that population by assuming H–W equilibrium and some reasonable sample size (cf. Svejgaard & Ryder, 1981, and our 'contrived' sample of the previous section).

Given the known heterogeneity of current urban populations, even within the less heterogeneous European countries, use of population control data culled, for example, from HLA workshop surveys, may alter the significance of calculated RR's. Although, in the examples given here the results are significant for both RR's and HRR's (Table 3), the '*p*-values' for significance differ by two-fold (for *DR3*) and 100-fold (for *DR4*), with the HRR's being more significant in each case. If less extreme samples were tested, careless choice of the control group could very well make the difference between statistical significance and non-significance (resulting in either a type I or a type II error).

Methods have previously been proposed for using sibship information to calculate 'risks'. For example, Clarke (1961) describes a method, attributed to C. A. B. Smith, for using sibships to test for a significant risk of duodenal ulcers to individuals of blood group O. The method used is somewhat different from that described here in that an observed and expected probability of being group O is assigned to the propositus in each sibship where the expected value depends on the makeup of the sibship. The significance is then based on a comparison of pooled observed and expected values over a set of sibships. This method does overcome the problem of heterogeneity but, because of the way the test is constructed, only a small part of the data can be used. In Clarke's example, therefore, the associations found when using the general population as a control were very much decreased when using Smith's sibship method. This does not seem to be the case using HRR's where the associations remain strong.

By using the two parental haplotypes not present in the single diseased individuals of the disease sample as the control 'sample', we are assured of having both samples from the same genetic population and, as was demonstrated above, this sample should represent a random sample of haplotype pairs (or 'individuals') from that population. Care must still be taken to ensure that the population chosen is genetically homogeneous, to the extent possible, but the task of obtaining an appropriate control is simplified.

If the disease is dominant rather than recessive, the HRR can still be used in the same way. Although it is not known whether the disease allele is present on the paternal haplotype ('*a*') or the maternal ('*c*') or perhaps on both, the other two parental haplotypes, *b* and *d*, will still represent random haplotypes from the underlying population, provided that the conditions mentioned for the recessive case obtain.

If a family is selected through more than one affected child, the situation is somewhat different. If the two affected sibs share the same two HLA haplotypes then the other two should still represent random haplotypes from the population. However, if they share fewer than two haplotypes, the situation is more complicated. Now three (or possibly four) haplotypes are known to carry the disease allele in the recessive case. If the disease is dominant, it is possible, but not certain, that a single shared haplotype carries the disease allele. If no haplotype is shared, it is not possible to define disease-carrying haplotypes with certainty. In such cases it would therefore be difficult to define a control sample of random haplotypes meeting the necessary criteria.

Two other points should be emphasized. If there is differential selection between genotypes at the susceptibility locus, (e.g. reduced fertility) a bias might be introduced such that the control haplotypes could no longer be considered a random population sample. Thus we require compliance with assumption (3) of our model to ensure the proper distribution of susceptibility alleles in the 'control' haplotypes.

Further, if the population from which the sample is drawn is genetically heterogeneous with respect to the disease, the HRR as well as the RR may be difficult to interpret as well as to use. In an extreme case a population might be made up of two ethnically distinct subpopulations that do not interbreed. Assume that the disease of interest occurs in only one of two such subpopulations. An estimate of the HRR would come entirely from a sample taken from the subpopulation where the disease is present and would be relevant only to that population (individuals in the other group having no risk, by definition). On the other hand, the RR would assign a risk over the entire population that would be too low for individuals in the susceptible part of the population and too high for individuals in the non-susceptible part.

We wish to thank Drs Jurg Ott, Neil Risch and C. A. B. Smith for helpful and constructive comments on an earlier draft of this paper.

This work was supported by NIH grant GM29177.

REFERENCES

Baur, M. P., Neugebauer, M. & Albert, E. D. (1984). Reference tables of two-locus haplotype frequencies for all MHC marker loci. In *Histocompatibility Testing* (eds. E. D. Albert, M. P. Baur and W. R. Mayr), pp. 677–755. Berlin: Springer-Verlag.

Bertrams, J. & Baur, M. P. (1984). Insulin-dependent diabetes mellitus. In *Histocompatibility Testing* (eds. E. D. Albert, M. P. Baur and W. R. Mayr), pp. 348–358. Berlin: Springer-Verlag.

Clarke, C. A. (1961). Blood Groups and Disease. *Progress in Medical Genetics* **1**, 81–119.

Curie-Cohen, M. (1981). HLA antigens and susceptibility to juvenile diabetes: do additive relative risks imply genetic heterogeneity? *Tissue Antigens* **17**, 136–148.

Falk, C. T., Mendell, N. R. & Rubinstein, P. (1983). Effect of population associations and reduced penetrance on observed and expected genotype frequencies in a simple genetic model: application to HLA and insulin dependent diabetes mellitus. *Ann. Hum. Genet.* **47**, 161–165.

Falk, C. T. (1984). A two-susceptibility-allele model for genetic diseases and associated marker loci: differences and similarities to a one-*s*-allele model. *Ann. Hum. Genet.* **48**, 87–95.

Rubinstein, P., Walker, M., Carpenter, C., Carrier, C., Krassner, J., Falk, C. & Ginsberg, F. (1981). Genetics of HLA disease associations. The use of the haplotype relative risk (HRR) and the "haplo-delta" (Dh) estimates in juvenile diabetes from three racial groups. *Human Immunology* **3**, 384 (Abstract).

Svejgaard, A. & Ryder, L. P. (1981). HLA genotype distribution and genetic models of insulin-dependent diabetes mellitus. *Ann. Hum. Genet.* **45**, 293–298.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Ann. Hum. Genet.* **19**, 251–253.

Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease [1]

Nathan Mantel *and* William Haenszel, *Biometry Branch, National Cancer Institute,*[2] *Bethesda, Maryland*

Summary

The role and limitations of retrospective investigations of factors possibly associated with the occurrence of a disease are discussed and their relationship to forward-type studies emphasized. Examples of situations in which misleading associations could arise through the use of inappropriate control groups are presented. The possibility of misleading associations may be minimized by controlling or matching on factors which could produce such associations; the statistical analysis will then be modified. Statistical methodology is presented for analyzing retrospective study data, including chi-square measures of statistical significance of the observed association between the disease and the factor under study, and measures for interpreting the association in terms of an increased relative risk of disease. An extension of the chi-square test to the situation where data are subclassified by factors controlled in the analysis is given. A summary relative risk formula, R, is presented and discussed in connection with the problem of weighting the individual subcategory relative risks according to their importance or their precision. Alternative relative-risk formulas, R_1, R_2, R_3, and R_4, which require the calculation of subcategory-adjusted proportions of the study factor among diseased persons and controls for the computation of relative risks, are discussed. While these latter formulas may be useful in many instances, they may be biased or inconsistent and are not, in fact, averages of the relative risks observed in the separate subcategories. Only the relative-risk formula, R, of those presented, can be viewed as such an average. The relationship of the matched-sample method to the subclassification approach is indicated. The statistical methodology presented is illustrated with examples from a study of women with epidermoid and undifferentiated pulmonary carcinoma.—**J. Nat. Cancer Inst. 22: 719–748, 1959.**

Introduction

A retrospective study of disease occurrence may be defined as one in which the determination of association of a disease with some factor is based on an unusually high or low frequency of that factor among diseased persons. This contrasts with a forward study in which one looks instead

[1] Received for publication November 6, 1958.

[2] National Institutes of Health, Public Health Service, U.S. Department of Health, Education, and Welfare.

for an unusually high or low occurrence of the disease among individuals possessing the factor in question. Each approach has its advantages. Among the desirable attributes of the retrospective study is the ability to yield results from presently collectible data, whereas the forward study usually requires future observation of individuals over an extended period (this is not always true; if the status of individuals can be determined as of some past date, the data for a forward study may already be at hand). The retrospective approach is also adapted to the limited resources of an individual investigator and places a premium on the formulation of hypotheses for testing, rather than on facilities for data collection. For especially rare diseases a retrospective study may be the only feasible approach, since the forward study may prove too expensive to consider and the study size required to obtain a respectable number of cases completely unmanageable.

In the absence of important biases in the study setting, the retrospective method could be regarded, according to sound statistical theory, as the study method of choice. This follows from the much reduced sample sizes required by this approach and may be illustrated by the following extreme example. If a disease attack rate of 10 per 100,000 among 50 percent of the population free of some factor were increased tenfold among the other half of the population subject to the factor, a retrospective study of 100 cases and 100 controls would, with high probability, reveal this significantly increased risk. On the other hand, a forward study covering 2,000 persons, half with and half without the factor, would almost certainly fail to detect a significant difference. For comparable ability to find the type of increased risk just indicated, a forward study would need to cover about 500 times as many individuals as the corresponding retrospective study. The disparity in the required number of persons to be studied could, of course, be reduced by lengthening the follow-up period for forward studies to increase the experience in terms of person-years observed. The larger sample size required for the forward study reflects principally the infrequent occurrence of the disease entity under investigation. In the example illustrated, uncovering 100 cases of disease in a forward study would require either 100,000 individuals with the factor or 1,000,000 without. For diseases with a higher probability of occurrence the disparity in required size between retrospective and forward studies would be progressively reduced.

The retrospective study might be looked upon as a natural extension of the practice of physicians since the time of Hippocrates, to take case histories as an aid to diagnosis. Its guise has varied with respect to the means of measuring the prevalence of the suspect factor among diseased persons and the criteria for determining unusual departures from normal experience. When an association is so marked, as in Percival Pott's observations on the representation of chimney sweeps among cases of scrotal cancer, no further quantitative data are required to perceive its significance.

The retrospective approach has often been employed in studies of communicable diseases, one illustration being Snow's observations (*1*) on a common water supply for cholera cases in an area served by several sources (there would have been no element of unusualness had there been but one water supply). When a disease is epidemic in a circumscribed locality, the disease-free population in the same area offers a natural contrast. The method may be used successfully for endemic diseases as well. Holmes, in reaching his conclusions on the communicable nature of puerperal fever (*2*), noted particularly that a large number of women with puerperal fever had been attended by the same physicians. In this context it should be emphasized that communicable disease investigations have often combined retrospective and forward study methods. For example, Snow supplemented his retrospective observations on water supply by a contrast of cholera rates among subscribers of the Southwark and Vaûxhall water company with the experience of persons served by the Lambeth water company within the same area.

When a disease occurs sporadically, or its occurrence is not confined to a well-defined group (such as women at childbirth), a choice of controls is not immediately evident. For cancer and other diseases characterized by high fatality rates, a study restricted to decedents might use persons dying from other causes as controls. Rigoni Stern adopted this technique in deducing the relationship of cancer of the breast and of the uterus to pregnancy history (*3*). Some contemporary studies have also used deaths from other causes as controls (*4*, *5*).

The present-day controlled retrospective studies of cancer date from the Lane-Claypon paper on breast cancer published in 1926 (*6*). This report is significant in setting forth procedures for selecting matched hospital controls and relating them to a consideration of study objectives. Retrospective techniques have since been applied in several investigations of cancer, including the following partial list of current references for a few primary sites: bladder (*7–10*), breast (*11–13*), cervix (*13–16*), larynx (*17*, *18*), leukemia (*19*), lung (*18*, *20–27*), and stomach (*13*, *28–30*).

Statisticians have been somewhat reluctant to discuss the analysis of data gathered by retrospective techniques, possibly because their training emphasizes the importance of defining a universe and specifying rules for counting events or drawing samples possessing certain properties. To them, proceeding from "effect to cause," with its consequent lack of specificity of a study population at risk, seems an unnatural approach. Certainly, the retrospective study raises some questions concerning the representative nature of the cases and controls in a given situation which cannot be completely satisfied by internal examination of any single set of data.

Only a few published papers have treated the statistical aspects of retrospective studies. Cornfield discussed the problem in terms of estimated measures of relative and absolute risks arising from contrasts of persons with and without specified characteristics (*31*). His paper was concerned with the simple situation of a homogeneous population of cases and controls, presumably alike in all characteristics except the one under investigation, which could be represented by a single contingency table. In a later contribution he handled the problem of controlling for other variables by adjusting the distribution of controls to the observed distribution of cases (*16*). Dorn briefly mentions retrospective studies with emphasis on such topics as sources of data, choice of controls, and validity of inferences (*32*).

This paper presents a method for computing relative risks for retrospective study contrasts, which controls for the effects of other variables by use of the basic statistical principle of subclassification of data. The related problem of significance testing is also considered. Since details of statistical treatment are conditioned by study objectives, data collection methods, choice of a control series, and the use of matched or unmatched controls, these topics are also discussed briefly.

Objectives

Retrospective studies are relatively inexpensive and can play a valuable role as scouting forays to uncover leads on hitherto unknown effects, which can then be explored further by other techniques. The effects may be novel and not suggested by existing data, as in the pioneer work on the association of smoking and lung cancer or the association of blood type and gastric cancer, or they may represent refinements of current knowledge. The latter category might include collection of lifetime residence and/or work histories to elaborate differences in incidence and mortality which appear when some diseases are classified by last place of residence or last occupation of the newly diagnosed case or decedent.

With diseases of low incidence the controlled retrospective study may be the only feasible approach. Here emphasis should be placed on assembling results from several studies. Before accepting a finding and offering an interpretation, scientific caution calls for ascertaining whether it can be reproduced by others and in other administrative settings having their own peculiar biases.

A primary goal is to reach the same conclusions in a retrospective study as would have been obtained from a forward study, if one had been done. Even when observations for a forward study have been collected, a supplementary retrospective approach to the same body of material may prove useful in collecting more data on points not covered in the original study design or in amplifying suggestive associations appearing in the initial forward-study results.

The findings of a retrospective study are necessarily in the form of statements about associations between diseases and factors, rather than about cause and effect relationships. This is due to the inability of the retrospective study to distinguish among the possible forms of association—cause and effect, association due to common causes, etc. Similar difficulties of interpretation arise in forward studies as well. A forward study, to avoid these difficulties, would need to be performed with the preciseness of a laboratory experiment. For example, such a study of associations with cigarette smoking would require that an investigator randomly assign his subjects in advance to the various smoking categories, rather than simply note the categories to which they belong. The inherent practical difficulties of such an enterprise are evident.

In addition to the failings shared with the forward study, the retrospective study is further exposed to misleading associations arising from the circumstances under which test and control subjects are obtained. The retrospective study picks up factors associated with becoming a diseased or a disease-free *subject*, rather than simply factors associated with presence or absence of the disease. The difficulties in this regard may be most pronounced when the study group represents a cross section of patients alive at any time (prevalence), including some who have been ill for a long period. Inclusion of the latter may lead to identification of items associated with the course of the illness, unrelated to increased or decreased risk of developing the disease. The theoretical point has been raised that factors conducive to longer survival of patients may be found in "prevalence" samples and interpreted erroneously as being associated with excess liability to the disease (*33*). Loopholes of this type are minimized when investigations are restricted to samples of newly diagnosed patients (incidence).

A partial remedy for these uncertainties lies in employing a conservative approach to interpretation of the associations observed. Recognizing the ease with which associations may be influenced by extraneous factors, the investigator may require not only that the measure of relative risk be significantly different from unity but also that it be importantly different. He may, for instance, require that the data indicate an increased relative risk for a characteristic of at least 50 percent, on the assumption that an excess of this magnitude would not arise from extraneous factors alone. However, the use of such conservative procedures emphasizes a corresponding need to pinpoint the disease entity under study. A strong relationship between a factor and a disease entity might fail to be revealed, if the entity was included in a larger, less well-defined, disease category. After the event from data now at hand, we know that a study of the association of cigarette smoking with epidermoid and undifferentiated pulmonary carcinoma is more revealing than an inquiry covering all histologic types of lung cancer.

Multiple Comparison Problem

The present-day retrospective study is usually concerned with investigating a variety of associations with a disease, little effort being involved in acquiring, within limits, added information from respondents. The results may be analyzed in a number of ways: the various factors may be investigated separately, without regard to the other factors; they may be investigated in conjunction with each other, a particular conjunction being considered a factor in its own right; or, more commonly, a factor may be tested with control for the presence or absence of other factors. Thus, if the role of cigarette smoking and coffee drinking in a given disease are under study, the possible comparisons include the relative risk of disease for individuals who both smoke and drink as opposed to all other persons, or as opposed to those who neither smoke, nor drink coffee. In addition, the relative risk associated with smoking might be obtained separately for drinkers and nondrinkers of coffee, with a weighted average of these two relative risks constituting still another item. Conversely, risks associated with coffee drinking, with adjustments for cigarette smoking, could be computed.

The potential comparisons arising from a comprehensive retrospective study can be large. Almost any reasonable level of statistical significance used to test a single contrast, when applied to a long series of contrasts, will, with a high degree of probability, result in some contrasts testing significant, even in the absence of any real associations. The usual prescription for coping with this multiple comparison problem—requiring individual comparisons to test significant at an extreme probability level to reduce the number of associations incorrectly asserted to be true—would result only in making real associations difficult to detect.

However, the multiple comparison problem exists only when inferences are to be drawn from a single set of data. If the purpose of the retrospective study is to uncover leads for fuller investigation, it becomes clear there is no real multiple significance testing problem—a single retrospective study does not yield conclusions, only leads. Also, the problem does not exist when several retrospective and other type studies are at hand, since the inferences will be based on a collation of evidence, the degree of agreement and reproducibility among studies, and their consistency with other types of available evidence, and not on the findings of a single study.

Nevertheless, it would be wise to employ testing procedures which do not lead to a superabundance of potential clues from any one study. This may be achieved by employing nominal significance levels in testing factors of primary interest incorporated into the design of an investigation and applying more stringent significance tests to comparisons of secondary interest or to comparisons suggested by the data. For the usual problem of multiple significance testing, this would be equivalent to allocating a large part of the desired risk of erroneous acceptance of an association as real to a small group of comparisons where fruitful results were anticipated, and parceling out the remainder of the available risk to the large bulk of comparisons of a more secondary nature. This minimizes the risk of diluting, through inclusion of many secondary comparisons, the chances for detecting an important primary effect.

Representative Nature of Data

The fundamental assumption underlying the analysis of retrospective data is that the assembled cases and controls are representative of the universe defined for investigation. This obligates the investigator not only to examine the data which are the end product but also to go behind the scenes and evaluate the forces which have channeled the material to his attention, including such items as local practices of referral to specialists and hospitals and the patient's condition and the effect of these items on the probability of diagnosis or hospital admission. We re-emphasize that this requires the exercise of judgment on the potential magnitude of biases and as to whether they could result in factors seeming to be related to a disease, in the absence of a real association of the factor with presence or absence of the disease. The danger of bias may be greatest in working with material from a single diagnostic source or institution.

Among the more important practical considerations affecting retrospective studies is that they are ordinarily designed to follow the line of least resistance in obtaining case and control histories. This means that cases and controls will often be hospital patients rather than persons in the general population outside hospitals. As a result, any factor which increases the probability that a diseased individual will be hospitalized for the disease may mistakenly be found to be associated with the disease. For example, Berkson (*34*) and White (*35*) have pointed out that positive association between two diseases, not present in the general population, may be produced when hospital admissions alone are studied, because persons with a combination of complaints are more likely to require hospital treatment. In theory, bias might also be produced in reverse manner, if the suspect factor diminished the probability of hospitalization for other diagnoses used as controls. The difficulties are not unique for hospital patients. Similar loopholes in interpretation may be advanced for any special groups used as sources of cases and controls.

However, a mere catalogue of biases arising from the possibly unrepresentative nature of a sample of cases and controls should not *ipso facto* invalidate any study findings. This is a substantive issue to be resolved on its merits for a specific investigation. Collateral evidence may provide information on the potential magnitude of bias and the size of spurious associations which could result. In some situations the difference between cases and controls may be so great that postulation of an unreasonably large bias would be required. Whether he consciously recognizes it or not, the investigator must always balance the risks confronting him and decide whether it is more important to detect an effect, when present, or to reject findings, when they may not reflect the true situation. If opportunities for further testing exist, one should not be too hasty in rejecting an association as an artifact arising from the method of data collection, and in foreclosing exploration of a potentially fruitful lead.

Because of the important role retrospective studies play in studies of human genetics, mention may be made of a bias frequently encountered in studies dealing with the familial distribution of diseases. A frequently used procedure takes a group of diagnosed cases for a disease in question and a group of controls and compares the prevalence of this disease among relatives of the probands and controls. The bias arises from the unrepresentative nature of the probands with respect to familial distribution and is known in other fields as "the problem of the index case" or "the effect of method of ascertainment." It has long been recognized that the characteristics for a random sample of families will differ from those for families to whom the investigator's attention has been directed because the family rosters include individuals selected for study on the basis of a specified attribute. For example, data on family size (number of children) obtained from siblings, rather than parents, are biased, since two or three potential index cases are present in the population for two- and three-child families as opposed to one for one-child families and none for childless couples. The analogy for disease occurrence is apparent. Families with two or three cases of the disease under study may have double or triple the probability of being represented by individuals in source material and having a representative selected as a proband than families with only one case. An appropriate analysis for this situation in studies of family size and birth order has been discussed by Greenwood and Yule (*36*), which takes account of the probability of family representation in proband data. Haenszel (*37*) has applied their correction to gastric-cancer data reported by Videbaek and Mosbech (*38*) and found the correction to reduce the originally reported fourfold excess of gastric cancer among relatives of probands, as compared to relatives of controls, to one of about 60 percent.

One remedy for the weakness of the retrospective approach to problems involving association of diseases and familial distribution would be to place greater reliance on forward observations of defined cohorts for data on these topics.

Controls

While easier accessibility to and lesser expense of hospital controls are important considerations, they should not deter one from collecting control data for a sample representing a more general population, if the latter are demonstrably superior. Some of the uncertainties about the superiority of hospital or general population controls arise from the need to maintain comparability in responses. The dependence of retrospective studies on comparability of responses from cases and controls cannot be overemphasized. When more accurate answers can be obtained from controls in a medical-care environment, the gain in comparability of responses for these controls could outweigh the other advantages to be derived from the more representative nature of general population controls. The difficulties may be illustrated by the experience with smoking histories. Hospital controls invariably yield a higher proportion of smokers for each sex than controls of comparable age drawn from the general population (*27*). Does this mean more complete smoking histories are collected in hospitals or does it imply that smokers have higher hospital admission rates? If the first alternative is correct, hospital controls are the appropriate choice for measuring the association of smoking history with a given disease. The second alternative calls for general population controls and in this situation the use of hospital controls yields underestimates of the degree of association.

Dual hospital and general population controls would have some merit. If control data from the two sources were in agreement, this would rule out some alternative interpretations of the findings. In the event of disagreement, its extent could be measured and alternate calculations made on the degree of association between an event and a suspect antecedent characteristic. Where the two sets of controls lead to substantially different results, a cautious and conservative interpretation is indicated.

Some topics, such as those bearing on sex practices and use of alcohol, may be amenable to study only within a clinical setting, and the collection of general population data on these items may prove impractical. The limitations of general population controls in this regard may have been overstressed, and empirical trials to test what information can be collected in household surveys should be encouraged instead of dismissing the possibility with no investigation whatsoever. Whelpton and Freedman, for example, have reported some success in collecting histories of contraceptive practices in interviews of a random sample of housewives (*39*).

When hospital controls are chosen, some precautions may be built into the study. Within limitations on the nature of controls imposed by a study hypothesis, controls drawn from a wide variety of diseases or admission diagnoses should be preferred. This permits examination of the distribution of the study characteristics among subgroups to check on internal consistency or variation among controls. This affords protection against two sources of error: *a*) attributing an association to the disease under investigation, when the effect is really linked to the diagnosis from which controls were drawn, and *b*) failure to detect an effect because both the study and control diseases are associated with the suspect factor. The latter is far from impossible. Both tuberculosis and bronchitis have exhibited association with smoking history and the use of one disease or the other as a control could easily lead to missing the association with smoking history. Similarly, patients with coronary artery disease would not constitute suitable controls for a study of the relationship of smoking and bladder cancer and *vice versa*, since the investigator would probably conclude that smoking was not related to either disease, when in truth it appears related to both. When there is definite evidence that two diseases are associated, for example, pernicious anemia and stomach cancer, the use of one as a control for the other is contraindicated, unless the study is specially designed to elucidate some aspects of the relationship.

It is always advantageous to include several items in a questionnaire for which general population data are available. This could be considered a partial substitute for dual hospital and general population controls. Disparity among cases, hospital controls, and general population controls on several general characteristics unrelated to the study hypothesis may be regarded as warning signals of the unrepresentative nature of the hospital cases and controls.

Where possible, interviews should be conducted without knowledge of the identity of cases and controls to guard against interviewer bias, although administrative reasons will often prevent attainment of "blind" interviews. In cooperative studies employing several interviewers, the magnitude of interviewer bias may be diminished, since it is unlikely that all interviewers will share the same bias in concert. In special circumstances, such as those prevailing at Roswell Park Memorial Institute, admissions may be interviewed before diagnosis, and hence before the identity of cases and controls is established. This feature requires a comprehensive, general purpose interview routinely administered to all admissions, which may restrict its use to publicly supported institutions diagnosing and treating neoplastic diseases or other specialized disease entities. Several epidemiological contiibutions for specific cancer sites have been based on the unique control data available from Roswell Park Memorial Institute (*9, 11, 12, 30, 40–43*), which are particularly valuable for collation with studies depending on more conventional sources of controls to evaluate interviewer bias and related issues.

Some patients interviewed as diagnosed cases will subsequently have their diagnoses changed. This may be turned to advantage. If scrutiny of the data for the erroneously diagnosed group reveals they had histories resembling those for the control rather than the case series, as Doll and Hill found in their study of smoking and lung cancer (*21*), this would constitute evidence against interviewer bias.

In investigations of a cancer site the association of a factor may often be restricted to a specific histologic type or a well-defined portion of an organ. The finding that epidermoid and undifferentiated pulmonary carcinoma is more strongly related to smoking history than adenocarcinoma of the lung is now well established. The range of explanations for the observed deficit of epidermoid carcinoma of the cervix in Jewish women as compared to other white women is greatly circumscribed by the presence of about equal numbers of adenocarcinoma of the corpus in both groups. When these finer diagnostic details or their significance are unknown to the interviewer, another check on interviewer bias is provided. Furthermore, the confirmation in repeated studies of an association limited to a specific histologic type or a detailed site will lend credence to an etiological interpretation of the association. Repeated confirmation is an essential element. Otherwise, a very specific association may be a reflection of the multiple comparison problem; if enough contrasts are created by fractionation of a single set of data, some apparently significant result is likely to appear. For this reason it would be desirable to reproduce such provocative results as Wynder's finding that use of alcohol was more strongly associated with cancer of the extrinsic larynx than of the intrinsic larynx (*18*), and Billington's report that prepyloric and cardiac neoplasms of the stomach were associated with blood group A and those located in the fundus with blood group O (*44*).

Discussion of matched controls in relation to the analysis and the computation of relative risks is deferred to a later section. One consideration on matched controls arising in the planning and development of a study should be mentioned here. Obviously, if the risk of disease changes with age an apparent association of the disease with other age-related factors may result. Other apparent associations with race, sex, nativity, etc., may arise in a similar manner. In devising rules for selecting controls, those factors known or strongly suspected to be related to disease occurrence should be taken into account if unbiased and more precise tests of the significance of the factors under investigation are desired. A sensible rule is to match those factors, such as age and sex, the effect of which may be conceded in advance and for which strong evidence is available from other sources, such as mortality data and morbidity surveys. When a factor is matched, however, it is eliminated as an independent study variable; it can be used only as a control on other factors. This suggests caution in the amount of matching attempted. If the effect of a factor is in doubt, the preferable strategy will be not to match but to control it in the statistical analysis. While the logical absurdity of attempting to measure an effect for a factor controlled by matching must be obvious, it is surprising how often investigators must be restrained from attempting this.

When a minimum of matching is involved, the importance of establishing, precisely and in advance, the method by which controls are selected for study increases. The rule should be rigid and unambiguous to avoid creating effects by subconscious selection and manipulation of controls. The problem is similar to that encountered in therapeutic trials where a protocol spelling out all the contingencies and actions to be taken in advance is, along with random assignment of cases and controls, the major bulwark against bias.

To reduce interview time and expense there are advantages in procedures for selecting controls which permit a case and the corresponding controls to be interviewed in a single session, particularly if travel to several institutions is involved. In practice, this favors selecting controls from a hospital patient census rather than from hospital admission lists. The difficulty with hospital admissions is that there is no guarantee that the controls will be available in the hospital at the time the diagnosed case is interviewed. This point seems more important than the fact that patients with diagnoses requiring long-term stays are overrepresented in a current hospital census (*45*). If the latter is an important issue, it may be handled in analysis through subclassification of controls by diagnosis.

Normally there will be little difficulty in reconciling these considerations into a harmonious set of rules. The items to be matched often lend themselves to a procedure for specifying controls. In a recent study on female lung cancer we found that the definition of two controls as the next older and the next younger women in the same hospital service, present on the day the case was interviewed, met the requirements just outlined (*27*). The controls were uniquely defined, the records establishing their identity were readily available on the service floor, interviews could be completed in one day, and a provision for balancing ages of cases and controls was incorporated. Simultaneous interviews of cases and controls may be more than an administrative convenience. If the prevalence of the associated factor is rapidly shifting over time, failure to control time of interview could obscure or exaggerate an association.

Some Statistical Tools

To progress further, questions on the representative nature of the case and control series must have been resolved affirmatively. With this condition in mind, let us suppose that a controlled retrospective study has been conducted and that the number of diseased cases, N_1, consists of A individuals with the factor being investigated and B free of the factor, while the number of controls, N_2, consists of C individuals with, and D individuals without the factor. Let $M_1 = A + C$, $M_2 = B + D$, $T = N_1 + N_2 = M_1 + M_2 = A + B + C + D$. What statistical evidence is there for the presence of an association and what is an appropriate measure of the strength of the association?

A commonly employed statistical test of association is the chi-square test on the difference between the cases and controls in the proportion of individuals having the factor under test. A corrected chi square may be calculated routinely as

$$(|AD-BC|-\tfrac{1}{2}T)^2T/N_1M_1N_2M_2$$

and tested as a chi square with 1 degree of freedom in the usual manner.

A suggested measure of the strength of the association of the disease with the factor is the apparent risk of the disease for those with the factor, relative to the risk for those without the factor. Consider that a population falls into the four possible categories and in the proportions indicated by the following table:

	With factor	Free of factor	Total
With disease	P_1	P_2	$P_1 + P_2$
Free of disease	P_3	P_4	$P_3 + P_4$
Total	$P_1 + P_3$	$P_2 + P_4$	1

The proportion of persons with the factor having the disease is $P_1/(P_1 + P_3)$, while the corresponding proportion for those free of the factor is $P_2/(P_2 + P_4)$. Relatively then, the risk of the disease for those with the factor is $P_1(P_2 + P_4)/P_2(P_1 + P_3)$. On a sampling basis this quantity may be estimated either by drawing a sample of the general population and estimating P_1, P_2, P_3, and P_4 therefrom or estimating $P_1/(P_1 + P_3)$ and $P_2/(P_2 + P_4)$ separately from samples of persons with, and persons free of, the factor.

It may be noted, however, that if the relative risk as defined equals unity, then the quantity P_1P_4/P_2P_3 will also equal unity. Further, for diseases of low incidence where the values for P_1 and P_2 are small in comparison with P_3 and P_4 it follows, as has been pointed out by Cornfield *(31)*, that P_1P_4/P_2P_3 is also a close approximation to the relative risk. This latter approximate relative risk can properly be estimated from the two sample approaches described or from samples drawn on a retrospective basis; that is, separate samples of persons with, and persons free of, the disease. The sample proportions of persons with, and free of, the factor in the retrospective approach provide estimates of $P_1/(P_1 + P_2)$ and of $P_2/(P_1 + P_2)$ from the sample having the disease and of $P_3/(P_3 + P_4)$ and of $P_4/(P_3 + P_4)$ from the disease-free sample. The estimate of P_1P_4/P_2P_3 is obtained by appropriate multiplication and division of these four quantities.

Whichever of the three methods of sampling is employed, the estimate of the approximate relative risk, P_1P_4/P_2P_3, reduces simply to AD/BC, where A, B, C, and D are defined in the manner stated in the first paragraph of this section. Also, the chi-square test of association given, which is essentially a test of whether or not the relative risk is unity, is equally applicable to all three sampling methods.

In the foregoing the two basic statistical tools of the epidemiologist for retrospective studies, the chi-square significance test and the measure of a relative risk, have been described for a relatively simple situation, one in which to all intents there is a single homogeneous population. The more complex situations confronting the epidemiologist in actual practice and the corresponding modifications in the statistical procedures will be presented.

Two other statistical problems may be noted here. One is the determination of how large a retrospective study to conduct. This depends on how sure we wish to be that the study will yield clear evidence that the relative risk is not unity, when it in fact differs from unity to some important degree. Application of this statistical technique requires reinterpreting a relative risk greater than unity into the corresponding difference between the diseased and the disease-free groups in the proportion of persons with the factor. For example, suppose an attack rate of 20 percent, given a normal rate of 10 percent, is worth uncovering. Suppose further that the factor associated with the increased disease rate affects 20 percent of the population. The population would then be distributed as follows:

	With factor	Free of factor	Total
With disease	$P_1=4\%$	$P_2=8\%$	12%
Free of disease	$P_3=16\%$	$P_4=72\%$	88%
Total	20%	80%	100%

The required retrospective study should be large enough to differentiate between a 33.3 percent $[P_1/(P_1 + P_2)]$ relative frequency of the factor among diseased individuals and an 18.2 percent $[P_3/(P_3+P_4)]$ relative frequency among disease-free individuals. The usual procedures for determining required sample sizes to differentiate between two binomial proportions are applicable in this situation.

While rigorous extension of this procedure to the more complex situations to be considered is not too simple, it can readily be adapted to secure approximations of the necessary study size. One might, for example, start by estimating the over-all required sample size following the procedure just indicated for differentiating between two sample proportions, assuming that cases and controls are homogeneous with respect to factors other than the one under investigation. Suppose on an over-all basis it is determined that the study should include $N_1 = 200$ disease cases and $N_2 = 200$ controls, but that the study data will be subclassified for purposes of analysis. Ignoring mathematical complications resulting from variations in binomial parameter values within individual subclassifications, we may interpret the above values of N_1 and N_2 as roughly meaning that the total information required for the study is $N_1N_2/(N_1 + N_2) = 100$. The objective should then be to assign values to N_{1i} and N_{2i} to obtain a total score of 100 for the cumulated information over all the subclassifications, $\Sigma N_{1i}N_{2i}/(N_{1i} + N_{2i})$, where N_{1i} and N_{2i} are the number of cases and controls in the *ith* subclassification.

This formulation of required total information brings out some aspects of retrospective study planning which are considered later in this paper. For instance, if any N_{1i} or N_{2i} is zero, no information is available from that particular category. Much of the benefit of a large N_{1i} (or N_{2i}) in any particular category is lost if the corresponding N_{2i} (or N_{1i}) is small. It is normally desirable to have N_{1i} and N_{2i} values commensurate with each other; for fixed totals, ΣN_{1i} and ΣN_{2i}, the total information in an investigation will be at a maximum if the degree of crossmatching is equal in all subclassifications with a constant case-control ratio of $\Sigma N_{1i}/\Sigma N_{2i}$. Maintaining a fixed case-control ratio among categories need not preclude assigning more cases and controls to specific categories. Larger numbers may be desired for categories of crucial interest to the study or for categories which represent greater segments of the population.

The information formula also reveals the limits for adjusting the relative numbers of diseased and control cases. It shows that if the number of controls (N_2) becomes indefinitely large, the required N_1 value can at most be reduced only by a factor of 2. Furthermore, this reduction in required diseased cases may be inappropriate if one wishes to obtain clear results for the separate subcategories.

The study size requirements suggested by the information formula may be seriously in error if the binomial parameters show excessive variation among subcategories. Ordinary precautions, however, should serve to keep the formula useful. In some situations it may be desirable to modify the information formula indicated above to reflect the contribution due to variation in the binomial parameters involved.

The second statistical procedure involves setting reasonable limits on the relative risk when it is in fact different from unity. For the homogeneous case considered, formulas for such limits have been published in (*46*). The chi-square test as stated is essentially a test of whether or not the confidence limits include unity. Extension of this procedure to more complex cases is fairly involved and depends primarily on the measure of relative risk adopted. In the absence of a clear justification for any single measure of over-all relative risk, the burden of extremely involved computation of confidence limits in such cases would not seem warranted. Instead, we feel that emphasis should be directed to obtaining an over-all measure of risk, coupled with an over-all test of statistical significance.

Statistical Procedures for Factor Control

A major problem in any epidemiological study is the avoidance of spurious associations. It has been remarked that where the risk of disease changes with age, apparent association of the disease with other age-related factors can result. However, there are appropriate statistical procedures for controlling those factors known or suspected to be related to disease occurrence. They serve not only to remove bias from the investigation but, in addition, can add to its precision.

Two simple procedures for obtaining factor control may first be mentioned. One is simply to restrict the investigation to individuals homogeneous on the factors to be controlled. For this situation the statistical procedures already outlined would be appropriate. The potential number of individuals available for such a study would, of course, be sharply restricted.

There is also the matching case method. A sample of N diseased individuals is drawn and the characteristics of each individual noted with respect to the control factors. Subsequently, a sample of N well individuals is drawn, with each individual matched on the control factors to one of the diseased individuals. The statistical procedures to be presented can be shown to cover the matched-sample approach as a special case, and a discussion of the analysis of such data will be given in that context. Some difficulties of the matched-sample study may be mentioned here. One is that when matching is made on a large number of factors, not even the fiction of a random sampling of control individuals can be maintained. Instead, one must be grateful for each matching control available. Another difficulty is that the method cannot be applied to factors under control, since diseased and control individuals are identical with respect to these factors. Conversely, factors under study in matched samples cannot themselves be controlled statistically. They can be analyzed separately or in particular conjunctions but cannot be employed as control factors.

An alternative to case matching is to draw independent samples of cases and controls, and adjust for other factors in the analysis. This approach requires simply the classification of individuals according to the various control and study factors desired, and an analysis for each separate subclassification as well as an appropriate summary analysis. Its success will depend on a reasonable degree of cross-matching between observations on diseased and control persons. In a small study various devices for reducing the number of subclassifications and for increasing the chances of cross-matching may be necessary, including a limit on the number of factors on which individuals are classified in any one analysis and the use of broad categories for any particular classification. Thus, a 10-year interval for age classification might permit a reasonable degree of cross-matching, whereas a 1-month interval would not.

The need for some degree of deliberate matching, even when the classification approach is employed, can be seen. If the disease under consideration occurs at advanced ages, little cross-matching would result if controls were selected from the general population. The remedy lies in deliberately selecting controls from the same age groups anticipated for persons with the disease, perhaps even matching one or more controls on age for each diseased person. This principle can be extended to matching on several control factors, *solely for the purpose of increasing the extent of cross-matching in the analysis.*

One of the subtle effects which can occur in a retrospective study, even with careful planning, may be pointed out. It can be shown, for instance, that within a given age interval the average age of individuals with cancer of certain sites will be greater than the average age of individuals from the general population in the same age interval. This can arise when incidence increases rapidly with age and may pose a serious problem with broad age intervals. This effect can be offset by close matching of cases and controls on age in drawing samples, even though they are classified by a broad age category in the analysis.

When a random sample of diseased and disease-free individuals is classified according to various control factors the distribution of the factor under study within the *ith* classification may be represented as follows:

	With factor	Free of factor	Total
With disease	A_i	B_i	N_{1i}
Free of disease	C_i	D_i	N_{2i}
Total	M_{1i}	M_{2i}	T_i

Within this subgroup the approximate relative risk associated with the disease may be written as A_iD_i/B_iC_i. One may compare the observed number of diseased persons having the factor, A_i, with its expectation under the hypothesis of a relative risk of unity, $E(A_i)=N_{1i}M_{1i}/T_i$. The discrepancy between A_i and $E(A_i)$ (which is also the discrepancy for any other cell within a 2 × 2 table) can be tested relative to its variance which, subject to the fixed marginal totals—N_{1i}, N_{2i}, M_{1i}, and M_{2i}—is given by $V(A_i) = N_{1i}N_{2i}M_{1i}M_{2i}/T_i^2(T_i-1)$. The corrected chi square with 1 degree of freedom $(|A_i-E(A_i)|-\frac{1}{2})^2/V(A_i)$ reduces in this case to $(|A_iD_i-B_iC_i|-\frac{1}{2}T_i)^2(T_i-1)/N_{1i}N_{2i}M_{1i}M_{2i}$. This formula for the variance of A_i is obtained as the variance of the binomial variable $N_1PQ(P = M_1/T$, $Q = M_2/T)$, multiplied by a finite population correction factor $(T-N_1)/(T-1) = N_2/(T-1)$. The earlier chi-square formula, which is ordinarily used, essentially employs a finite population correction factor of N_2/T.

There is thus a difference between the two chi-square formulas of a factor of $(T-1)/T$ which, though trivial for any single significance test with respectably large T, can become important in the over-all significance test. It is with the latter formula, just presented, that chi square is computed as the ratio of the square of a deviation from its expected value to its variance.

The adjustment for control factors is at this point resolved for the resulting separate subclassifications. The problem of over-all measures of relative risk and statistical significance still remains. A reasonable over-all significance test which has power for alternative hypotheses, where there is a consistent association in the same direction over the various subclassifications between the disease and a study factor, is provided by relating the summation of the discrepancy between observation and expectation to its variance. The corrected chi square with 1 degree of freedom then becomes $(|\Sigma A_i - \Sigma E(A_i)| - \frac{1}{2})^2/\Sigma V(A_i)$ where $E(A_i)$ and $V(A_i)$ are defined as above.

The specification of a summary estimate of the relative risk associated with a factor is not so readily resolved as that for an over-all significance test, and involves consideration of alternate approaches to a weighted average of the approximate relative risks for each subclassification (A_iD_i/B_iC_i). If one could assume that the increased relative risk associated with a factor was constant over all subclassifications, the estimation problem would reduce to weighting the several subclassification estimates according to their respective precisions. The complex maximum likelihood iterative procedure necessary for obtaining such a weighted estimate would seem to be unjustified, since the assumption of a constant relative risk can be discarded as usually untenable.

Another possible criterion for obtaining a summary estimate of relative risk would involve weighting the risks for subclassification by "importance." A twofold increase of a large risk is more important than a twofold increase of a small risk. An increased risk for a large group is more important than one for a small group. An increased risk for young individuals may be more important than for older individuals with a shorter life expectation. Difficulties arise in attempts to weight relative risk by measures of importance. For one, the necessary information on importance, in terms of the size of the populations affected or in terms of the absolute level of rates prevailing in the subgroups, is generally not contained within the scope of the investigation. A problem in definition of the precise terms of the weighted comparison also appears. Does one want to adjust the risks of disease among persons with the factor to the distribution of the population without the factor, or *vice versa*, or adjust the risks for the populations with and without the factor to a combined standard population? These procedures, and the different phrasing of the comparisons which they entail, could yield different answers. If only a small proportion of the population with the factor was in a subcategory with a high relative risk, while most of the factor-free population fell into this subcategory, and in other categories the relative risk associated with the factor was less than unity, the factor would appear to exert a protective influence under one set of weights but a harmful effect under the other.

Published instances of summary relative risks do not fall clearly into either of the two categories—weighting by precision or weighting by importance. They do follow an approach usually employed in age-adjusting mortality data. Since the relative risk for a single 2 × 2 table can be obtained from the incidence of the factor among diseased and well individuals, the problem would appear translatable into terms of obtaining over-all, category-adjusted incidence figures. Direct or indirect methods of adjustment can be used, employing as a standard of reference the frequency distribution or rates corresponding to the sample of diseased persons, of controls, or the diseased persons and controls combined.

While such adjustment procedures provide weighting by importance in their customary application to mortality rates, this is not so in the relative risk situation. This may be illustrated in the following extreme example. Suppose that in each of two subcategories the approximate relative risk for a contrast between the presence and absence of a factor is about 5, which arises in the first subcategory from contrasting percentages of 1 and 5, and in the second subcategory from contrasting percentages of 95 and 99. If these percentages were based on equal numbers of individuals, all methods of category adjusting would yield contrasting adjusted summary percentages of 46 and 52, and a resultant relative risk of slightly less than 1.3. Some other approach for obtaining category-adjusted relative risks would seem desirable. However, to the extent that such extreme situations are not encountered in actual practice, results based on these more conventional adjustment procedures will not be grossly in error.

A suggested compromise formula for over-all relative risk is given by $R = \Sigma(A_iD_i/T_i)/\Sigma(B_iC_i/T_i)$. As a weighted average of relative risks this formula would, in the illustration given, yield the over-all relative risk of 5 found in each of the two subcategories. The weights are of the order $N_{1i}N_{2i}/(N_{1i} + N_{2i})$ and as such can be considered to weight approximately according to the precision of the relative risks for each subcategory. The weights can also be regarded as providing a reasonable weighting by importance.

An interesting property of this summary relative risk formula is that it equals unity only when $\Sigma A_i = \Sigma E(A_i)$ and hence the corresponding chi square is zero. From the fact that $A_i - E(A_i) = (A_iD_i - B_iC_i)/T_i$, it follows that when $\Sigma A_i = \Sigma E(A_i)$, $\Sigma A_iD_i/T_i$ will equal $\Sigma B_iC_i/T_i$, chi square will be zero, and R will be unity. The chi-square significance test can thus be construed as a significance test of the departure of R from unity.

Of some other procedures for measuring over-all relative risks, the one following also has the interesting property of being equal to unity when $\Sigma(A_i) = \Sigma E(A_i)$ and therefore subject to the chi-square test:

$$R_1 = \frac{\Sigma A_i \Sigma D_i}{\Sigma B_i \Sigma C_i} \Bigg/ \frac{\Sigma E(A_i) \Sigma E(D_i)}{\Sigma E(B_i) \Sigma E(C_i)} \text{ where } E(A_i) = N_{1i}M_{1i}/T_i,\ E(B_i)$$

$$= N_{1i}M_{2i}/T_i,\ E(C_i) = N_{2i}M_{1i}/T_i, \text{ and } E(D_i) = N_{2i}M_{2i}/T_i.$$

In this formula the numerator represents the crude value for the relative risk, which would result from pooling the data into one table and ignoring all subclassification on other factors. The denominator represents the crude value for relative risk, which would have resulted from pooling in the situation where all relative risks within each subclassification were exactly unity. Readers familiar with the "indirect" method of computing standardized mortality ratios will recognize an analogy between the "indirect" method and the above procedure.

The estimator R_1 can be seen to have a bias toward unity. One reason is covered by the illustration which indicated that adjusted percentages (or frequencies) do not yield an appropriate adjusted relative risk. In addition, when either cases or controls have little representation in a subcategory, there will be lack of cross-matching and little information about relative risk, and the observed cell frequencies and their expectations will be numerically close. Such results will,in the process of summation used by the estimator, tend to force its value toward unity. This weakness will not be too important if the degree of cross-matching is roughly equal in the various subclassifications—an optimum goal one would normally attempt to achieve. The bias will become more pronounced as the number of control factors increases and as the prospects for good cross-matching become poorer.

We used the estimator R_1 in a recent paper (*27*), knowing its potential weaknesses. This was done to present results more nearly comparable with those reported by other investigators using similarly biased estimators. One set of results from this paper on lung cancer among women illustrates the conservative behavior of estimator R_1 compared with R, as additional factors are controlled. The relative risk (R_1) for epidermoid and undifferentiated pulmonary carcinoma associated with smoking more than one pack of cigarettes daily as compared to nonsmokers decreased from 7.1 (controlled for age) to 5.6 (controlled for age and coffee consumption). The corresponding figures, with R as a measure of relative risk, were 9.7 and 9.9.

Computational procedures for R and R_1 are presented in table 1, drawing on material comparing smoking histories of women diagnosed as cases of epidermoid and undifferentiated pulmonary carcinoma with those of female controls. For simplicity in presentation only two smoking levels are considered—nonsmokers and smokers of more than one pack of cigarettes daily. An extension of the significance testing procedures to the case of study factors at more than two levels is discussed later. The control factors are age and occupation. The basic data are given in the first 9 columns. Columns 10 and 11 carry the derivative calculations required for R. Columns 12 and 13 are used in the computation for R_1 and for the variance estimate in column 14—the latter being needed for the chi-square test. Only columns 1 to 10, 12, and 14 would be necessary to compute chi square, R and R_1. Column 13 is not essential for the computation of $E(D)$ but simplifies computation of $V(A)$, while providing a check on $E(A)$.. Column 11 serves as a check on 10 and 12. A system of checks and computations is outlined at the bottom of table 1. Not all the computations shown would ordinarily be necessary for an analysis.

The corrected chi-square value of 30.66 (1 degree of freedom) would indicate a highly significant association between epidermoid and undifferentiated pulmonary carcinoma and cigarette smoking in women, after adjusting for possible effects connected with age or occupation. The

TABLE 1.—*Illustrative computations for chi square and for summary measures of undifferentiated pulmonary carcinoma*

Group		Epidermoid-undifferentiated pulmonary carcinoma			Controls			Cases and controls		
		1 + Pack cigarettes daily	Nonsmokers	Total	1 + Pack cigarettes daily	Nonsmokers	Total	1 + Pack cigarettes daily	Nonsmokers	Total
		A (1)	B (2)	N_1 (3)	C (4)	D (5)	N_2 (6)	M_1 (7)	M_2 (8)	T (9)
Housewives	under age 45	0	2	2	0	7	7	0	9	9
	45–54	2	5	7	1	24	25	3	29	32
	55–64	3	6	9	0	49	49	3	55	58
	65 and over	0	11	11	0	42	42	0	53	53
White-collar workers	under age 45	3	0	3	2	6	8	5	6	11
	45–54	2	2	4	2	18	20	4	20	24
	55–64	2	4	6	2	23	25	4	27	31
	65 and over	0	6	6	1	11	12	1	17	18
Other occupations	under age 45	1	0	1	3	10	13	4	10	14
	45–54	4	1	5	1	12	13	5	13	18
	55–64	0	6	6	1	19	20	1	25	26
	65 and over	1	3	4	0	15	15	1	18	19
Total		18	46	64	13	236	249	31	282	313

Checks: Total discrepancy, $Y, = \Sigma A - \Sigma E(A) = \Sigma(1) - \Sigma(12) = 11.625$

$= \Sigma D - \Sigma E(D) = \Sigma(5) - \Sigma(13) = 11.625$

$= \Sigma(AD/T) - \Sigma(BC/T) = \Sigma(10) - \Sigma(11) = 11.625$

$\Sigma(15) + \Sigma(16) = 64.000; \Sigma(3) = 64$

$\Sigma(17) + \Sigma(18) = 249.000; \Sigma(6) = 249$

Derivative computations: $\Sigma E(B) = \Sigma(2) + Y = 57.625$

$\Sigma E(C) = \Sigma(4) + Y = 24.625$

$\Sigma(AT/N_1) = \Sigma(1) + \Sigma(17) = 94.960$

$\Sigma(BT/N_1) = \Sigma(2) + \Sigma(18) = 218.040$

$\Sigma(CT/N_2) = \Sigma(4) + \Sigma(15) = 16.325$

$\Sigma(DT/N_2) = \Sigma(5) + \Sigma(16) = 296.675$

value of R implies that the risk of these cancers is 10.7 times as great for women currently smoking in excess of 1 pack a day than for women who never used cigarettes. The value of R_1, 7.05, is almost identical with the crude relative risk, 7.10, which results from pooling the data with no attention to the control factors. The difference from the published R_1 value of 6.3 in (*27*) arises from the exclusion in the illustrative example, of data for women currently smoking 1 pack a day or less and for occasional or discontinued smokers.

The computation of three other summary estimates of relative risk is also outlined in table 1. The additional derivative computations required for this purpose appear in columns 15 to 18. All three estimates are based on a direct method of category adjustment, that is, the use of a standard distribution to which both the case and control distributions are *relative risk* (R, R_1, R_2, R_3, *and* R_4) *relating to the association of epidermoid and in women with smoking history*

Derivative computations								
$\frac{AD}{T}$	$\frac{BC}{T}$	E(A)	E(D)	V(A)	$\frac{N_1C}{N_2}$	$\frac{N_1D}{N_2}$	$\frac{N_2A}{N_1}$	$\frac{N_2B}{N_1}$
$\frac{(1)(5)}{(9)}$	$\frac{(2)(4)}{(9)}$	$\frac{(3)(7)}{(9)}$	$\frac{(6)(8)}{(9)}$	$\frac{(12)(13)}{(9)-1.0}$	$\frac{(3)(4)}{(6)}$	$\frac{(3)(5)}{(6)}$	$\frac{(1)(6)}{(3)}$	$\frac{(2)(6)}{(3)}$
(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
0	0	0	7. 000	0	0	2. 000	0	7. 000
1. 500	0. 156	0. 656	22. 656	0. 480	0. 280	6. 720	7. 143	17. 857
2. 534	0	0. 466	46. 466	0. 380	0	9. 000	16. 333	32. 667
0	0	0	42. 000	0	0	11. 000	0	42. 000
1. 636	0	1. 364	4. 364	0. 595	0. 750	2. 250	8. 000	0
1. 500	0. 167	0. 667	16. 667	0. 483	0. 400	3. 600	10. 000	10. 000
1. 484	0. 258	0. 774	21. 774	0. 562	0. 480	5. 520	8. 333	16. 667
0	0. 333	0. 333	11. 333	0. 222	0. 500	5. 500	0	12. 000
0. 714	0	0. 286	9. 286	0. 204	0. 231	. 769	13. 000	0
2. 667	0. 056	1. 389	9. 389	0. 767	0. 385	4. 615	10. 400	2. 600
0	0. 231	0. 231	19. 231	0. 178	0. 300	5. 700	0	20. 000
0. 790	0	0. 211	14. 211	0. 166	0	4. 000	3. 750	11. 250
12. 825	1. 201	6. 375	224. 375	4. 036	3. 325	60. 675	76. 960	172. 040

Chi-square: $X^2 = (|\text{discrepancy}| - 0.5)^2/\Sigma V(A) = (|Y| - 0.5)^2/\Sigma(14) = 30.66$
Relative risk: $R = \Sigma(AD/T)/\Sigma(BC/T) = \Sigma(10)/\Sigma(11) = 10.68$

$$R_1 \begin{cases} \text{crude relative risk, } r = \Sigma A \Sigma D/\Sigma B \Sigma C = \Sigma(1)\Sigma(5)/\Sigma(2)\Sigma(4) = 7.10 \\ \text{adjustment factor, } f = \Sigma E(A)\Sigma E(D)/\Sigma E(B)\Sigma E(C) = \Sigma(12\Sigma(13)/\Sigma E(B)\Sigma E(C) \\ \quad = 1.0081 \\ R_1 = r/f = 7.05 \end{cases}$$

$R_2 = \Sigma A \Sigma(N_1D/N_2)/\Sigma B \Sigma(N_1C/N_2) = \Sigma(1)\Sigma(16)/\Sigma(2)\Sigma(15) = 7.14$
$R_3 = \Sigma(N_2A/N_1)\Sigma D/\Sigma(N_2B/N_1)\Sigma C = \Sigma(5)\Sigma(17)/\Sigma(4)\Sigma(18) = 8.12$
$R_4 = \Sigma(AT/N_1)\Sigma(DT/N_2)/\Sigma(BT/N_1)\Sigma(CT/N_2) = 7.91$

Note: Figures shown are rounded from those actually calculated and consequently are not fully consistent. Column totals and figures shown do not necessarily agree.

adjusted. If the distribution of diseased cases is taken as the standard distribution to which the controls are adjusted, the estimator becomes

$$R_2 = \frac{\Sigma A_i \Sigma\left(D_i \times \frac{N_{1i}}{N_{2i}}\right)}{\Sigma B_i \Sigma\left(C_i \times \frac{N_{1i}}{N_{2i}}\right)}.$$

Estimator R_2 was used by Wynder *et al.* in a study of the association of cervical cancer in women with circumcision status of sex partners (*16*). The merit of employing the cervical cancer case-distribution as the standard presumably rests on the fact that this distribution at least would be well defined by the study.

If the distribution of control cases is taken as standard the estimator becomes

$$R_3 = \frac{\Sigma\left(A_i \times \frac{N_{2i}}{N_{1i}}\right)\Sigma D_i}{\Sigma\left(B_i \times \frac{N_{2i}}{N_{1i}}\right)\Sigma C_i}.$$

If the combined distribution is taken as standard the estimator becomes

$$R_4 = \frac{\Sigma\left(A_i \times \frac{T_i}{N_{1i}}\right)\Sigma\left(D_i \times \frac{T_i}{N_{2i}}\right)}{\Sigma\left(B_i \times \frac{T_i}{N_{1i}}\right)\Sigma\left(C_i \times \frac{T_i}{N_{2i}}\right)}.$$

If any N_{1i} or N_{2i} should equal zero, the estimator R_4 would not be defined. R_2 is not defined for any zero-valued N_{2i}, and R_3 is not defined for any zero-valued N_{1i}. In these instances it would be necessary to exclude the zero-frequency categories to define the estimators. The estimator R_1 retains these categories at the expense of greater bias toward unity. The estimator R gives such categories zero weight, since they contain no information about relative risk. The chi-square significance test gives no weight to these categories.

While R_4 is clearly a direct adjusted estimate of relative risk employing the combined distribution as standard, R_2 and R_3 may be viewed alternatively as either direct or indirect adjusted estimates. The same estimates will result if a direct adjustment is made using the distribution of cases as standard, or an indirect adjustment is made using the factor incidence rates for controls as the standard rates.

It may be noted that in the example used, the values for R_2, R_3, and R_4 (7.14, 8.12, and 7.91, respectively) were roughly comparable to R_1, and all were smaller than R. The example was selected because all the N_{1i} and N_{2i} values were non-zero, so that the values of R_2, R_3, and R_4 were all defined.

The over-all relative risk estimates are averages and as averages may conceal substantial variation in the magnitudes of the relative risk among subgroups. Ordinarily, the individual subcategory data should be examined, paying special attention to relative risks based on reasonably large sample sizes. This will provide protection against the potential deficiencies of any particular summary relative risk formula employed. The over-all chi-square significance test in any case will remain appropriate for detecting any strong general tendency for the risk of disease to be associated with the presence or absence of the test factor.

The Matched-Sample Study

The matched-sample study previously described can be considered a special case of the classification procedure with the number of classifications equal to the number of pairs of individuals. The status of pairs of well and diseased individuals classified with respect to the presence or absence of the suspect factor in each individual will be represented as

F, G, H, or J in the following fourfold table. The meanings attached to the marginal totals A, B, C, and D are the same as those in the first schematic representation.

	Diseased individuals		
Well individuals	With factor	Free of factor	Total
With factor	F	G	C
Free of factor	H	J	D
Total	A	B	N

In the absence of association between the disease and the factor, we expect the same number of individuals with the factor to appear among both diseased and well individuals; that is, we expect $A(=F + H)$ to equal $C(=F + G)$. This can occur only when $G = H$ and the statistical test is simply whether or not G differs significantly from 50 percent of $G + H$. G is tested as a binomial variable with parameter $\frac{1}{2}$, $G + H$ being the number of cases. G thus has expectation $\frac{1}{2}(G + H)$, variance $\frac{1}{4}(G + H)$ and the corrected chi square with 1 degree of freedom can readily be shown to reduce to $(|G - H| - 1)^2/(G + H)$.

Treating the data as consisting of N classifications each with $N_{1i} = N_{2i} = 1$, $T_i = 2$ and applying the previously described procedures will lead to the same value of chi square. For F of the N classifications, $A_i = 1$, $M_{1i} = 2$, $M_{2i} = 0$, $E(A_i) = 1$, $V(A_i) = 0$; for G classifications $A_i = 0$, $M_{1i} = M_{2i} = 1$, $E(A_i) = \frac{1}{2}$, $V(A_i) = \frac{1}{4}$; for H classifications $A_i = 1$, $M_{1i} = M_{2i} = 1$, $E(A_i) = \frac{1}{2}$, $V(A_i) = \frac{1}{4}$; and for J classifications, $A_i = 0$, $M_{1i} = 0$, $M_{2i} = 2$, $E(A_i) = 0$, $V(A_i) = 0$. Thus, $\Sigma A_i = F + H$, $\Sigma E(A_i) = F + \frac{1}{2}(G + H)$, $\Sigma V(A_i) = \frac{1}{4}(G + H)$, and the resultant corrected chi square can again be seen to be $(|G - H| - 1)^2/(G + H)$.

It is of interest to observe that the summary chi-square formula is appropriate in the matched-sample case, even though the frequencies for each of the separate subclassifications are small. Its appropriateness, despite the small frequencies, stems from the fact that it is a test on a summation of random variables, A_i, and thus tends to approach normality rapidly, making the chi-square test valid, even though the individual A_i's are not normally distributed. This property of the chi-square formula applies in the general classification as well as the matched-sample situation. Only substantial lack of cross-matching in the general case would tend to make the chi-square test invalid. It is also essential, of course, that there be some appreciable variation in the presence or absence of the factor under study.

It should be noted that in the matched-sample study with $T_i = 2$ for each of the N pairs of individuals, the variances of the A_i's would have been understated by a factor of 2, had $T - 1$ been replaced by T in the variance formulas. The usual formula for chi square does essentially make this replacement, but it is usually of little consequence if T is of any reasonable magnitude. The formulas for relative risk in the matched-sample study reduce simply to the following: $R = H/G$; $R_1 = R_2 = R_3 = R_4 = AD/BC$.

Study Factors at More Than Two Levels

The preceding discussion on the analysis of retrospective data has been in terms of the test factor under study taking only two values. This framework has sufficed for discussion of the underlying statistical ideas and issues. In practice, the study factor will frequently take on more than two, perhaps many, potential values. When the number of study factor values is large, grouping can reduce them to manageable proportions.

The need to consider only a limited number of classes for the study factor stems from the fact that, when an association is anticipated, most of the significant information about the association will come from the results for the more extreme values of the study factor. While it is efficient to concentrate attention on the test factor classes expected to show the greatest differences in association with the disease, it is also profitable to consider intermediate values for the test factor to seek evidence for a consistent pattern of association. For example, in table 1, a highly significant difference between nonsmokers and women currently smoking more than 1 pack of cigarettes daily was illustrated. Inclusion of data for smokers of 1 pack or less a day showing results intermediate between the other classes would have added little, if anything, to the statistical significance of the results, and might actually lower it, if one made an over-all test of the differences among the three smoking classes. However, the observation that the intermediate smoking class does, in fact, show an intermediate relative risk contributes to an orderly pattern and increases our confidence in the conclusions suggested by the data for the remaining two classes.

For any two particular test-factor levels, the relative risk for one over the other may be calculated using only the data pertaining to those two levels or by using the results for all test levels. In the formulas previously given for R, R_1, R_2, R_3, and R_4, the difference between the two calculating procedures is simply one of setting the values of N_{1i}, N_{2i}, and $T_i = N_{1i} + N_{2i}$ in terms of number of cases and controls occurring at the two study-factor levels only, or defining them in terms of total number of cases and controls in the entire study. When total cases and controls are used in defining N_{1i}, N_{2i}, and T_i, it can be shown that for R_1, R_2, R_3, and R_4 the various relative risks will be internally consistent with each other. If the relative risk for the first level is twice that for the second level, which in turn is twice that for the third level, then the relative risk for the first level will be four times that of the third. These exact relationships do not hold for R as an estimator of relative risk, and a somewhat sophisticated extension of the formula for R would be required to secure this property.

The problem of obtaining a summary chi square when the study factor is at more than two levels is complicated by the fact that the deviations from expectation at the various study-factor levels are intercorrelated. When there are but two levels, the two deviations will have perfect negative correlation, and attention need be directed to only one of the deviations. Irrespective of the number of levels, at any one level the deviation from expectation among diseased persons will be equal, but opposite in sign, to the deviation from expectation among controls, so that attention can be confined to the deviations for diseased persons.

The problem can be stated as one of reducing a set of correlated deviations into a summary chi square. Table 2 applies this process for obtaining a summary chi square to the study of the association of epidermoid and undifferentiated pulmonary carcinoma in women and maximum cigarette-smoking rate, classified into three levels, after adjustment for age and occupation.

The general expressions for the expectations and variances of the number of cases at a particular test-factor level are given in the lower right section of table 2. Also shown is the expression for the covariance between the number of cases at two different test-factor levels. Since the total of all the deviations is zero, one would in general need the variances of, and covariances between, the number of cases at all but one of the levels. The number of covariance terms will rise sharply as the number of test levels are increased. At 3 test levels, there are 2 variance terms and 1 covariance term, while at 10 test levels, there would be 9 variances and 36 covariance terms of interest.

For the general case the burden of computation could be heavy. After all the necessary computation for the deviations, their variances and covariances, there would still remain the problem of converting these, presumably by matrix methods, into a surmary chi square. Since the retrospective problem will normally involve only a limited number of test-factor levels, precise procedures will be given only for the three-level situation, and approximate procedures outlined for the general case.

The exact computation procedure for the three-level case is detailed in table 2. Lines (1), (2), and (4) show the total observed and expected frequencies and variances of the number of cases (and controls) at each of the three smoking-rate levels, after adjusting for age and occupation. These are the summary totals over each subclassification obtained by application of the formulas appearing in table 2.

Lines (5) and (6) give the chi squares corresponding to the total deviation from expectation at each of the smoking-rate levels. The chi squares in line (5) are corrected for continuity. They relate to the difference of the particular level to which they apply, from the two other levels combined. Following the usual practice of making no continuity corrections when chi squares with more than 1 degree of freedom are under consideration, line (6) shows the uncorrected chi squares.

The computing procedure of table 2 takes advantage of the fact that, since the sum of the deviations from expectation is zero, the variance of the third deviation must equal the sum of the other two variances plus twice the covariance for the first two deviations. The covariance of the first two deviations is readily obtained as illustrated and is used in calculating the summary chi square. The summary chi square is obtained as the sum of squares of two orthogonal deviates, with each

TABLE 2.—*Illustrative computation of summary chi square, when there are 3 levels for study factor. The data relate to the association of epidermoid and undifferentiated pulmonary carcinoma in women with smoking history*

	1+ Pack cigarettes daily			1 Pack or less of cigarettes daily			Occasional or nonsmokers			Total		
	Epidermoid-undifferentiated pulmonary carcinoma	Controls	Total (ΣM_1)	Epidermoid-undifferentiated pulmonary carcinoma	Controls	Total (ΣM_2)	Epidermoid-undifferentiated pulmonary carcinoma	Controls	Total (ΣM_3)	Epidermoid-undifferentiated pulmonary carcinoma (ΣN_1)	Controls (ΣN_2)	Total (ΣT)
(1) Total observed frequencies	19	17	36	32	71	103	51	251	302	102	339	441
(2) Total expected frequencies, adjusted for age and occupation	9. 09	26. 91	36	23. 76	79. 24	103	69. 15	232. 85	302	102	339	441
(3) Total deviation from expectation (1)−(2)	$+9.91 = Y_1$			$+8.24 = Y_2$			$-18.15 = Y_3$			For the general situation the total expected case frequency at the *jth* level of a test factor is $\sum_i N_{1i}M_{ji}/T_i$ The variance of the total case frequency is $V_j=\sum_i\frac{N_{1i}N_{2i}M_{ji}(T_i-M_{ji})}{T_i^2(T_i-1)}$ The covariance of the total case frequencies at test levels *j* and *k* is $-\sum_i\frac{N_{1i}N_{2i}M_{ji}M_{ki}}{T_i^2(T_i-1)}$ The index of summation, *i*, represents the various subclassifications into which the results are divided For 3 test levels only, since $Y_3=-(Y_1+Y_2)$, it follows that $V_3=V_1+V_2+2$ Covariance (Y_1,Y_2)		
(4) Variance of total observed frequencies, subject to fixed marginal totals in each age and occupation group	$5.9163 = V_1$			$12.2900 = V_2$			$14.0723 = V_3$					
(5) Individual corrected chi squares $(\|Y\|-0.5)^2/V$	$14.97 = \chi^2_{1c}$			$4.88 = \chi^2_{2c}$			$22.15 = \chi^2_{3c}$					
(6) Individual uncorrected chi squares Y^2/V	$16.60 = \chi^2_1$			$5.53 = \chi^2_2$			$23.42 = \chi^2_3$					
(7) Covariance (Y_1,Y_2) $(V_3-V_1-V_2)/2$				-2.0670								
(8) Adjusted Y_2 $Y_2-(7)Y_1/V_1$				11.70								
(9) Adjusted V_2 $V_2-(7)^2/V_1$				11.5678								
(10) Adjusted χ^2_2 $(8)^2/(9)$				$11.83 = \chi^2_2$(ad.)								
(11) Summary chi square (2 degrees of freedom) $\chi^2_1 + \chi^2_2$(ad.)				$16.60 + 11.83 = 28.43$								

square adjusted for its own variance. The first deviate squared is simply the uncorrected chi square at the first level in line (6)—the variance of the deviate remaining as initially calculated. The second deviate is the deviation at the second level adjusted for its correlation with the first deviation [adjusted $Y_2 = Y_2 - b_{21}Y_1$; b_{21} = covariance (Y_1,Y_2)/variance Y_1)]. The variance of the adjusted second deviate is the initial value reduced by that portion of the variation accounted for by the first deviation [Var. (adjusted Y_2) = variance Y_2—covariance$^2(Y_1,Y_2)$/variance Y_1)].

In the present instance the summary chi square with 2 degrees of freedom is 28.43 [line (11)]. This presumably is close to the chi square with 1 degree of freedom which would have obtained had only the two most extreme smoking classes been compared. If one examines the individual uncorrected chi squares [line (6)], their total is found to be 45.55, the maximum individual figure being 23.42. *It will necessarily be true that the summary chi-square value will lie between the largest of the three chi squares and their total. At almost any reasonable probability level these limits would be sufficient to establish statistical significance without further calculation.* In our companion paper (*27*) this rule sufficed in almost all instances to separate the significant from the nonsignificant results.

Comments on Extensions to More Than Three Factors

Two procedures can be suggested for getting approximate summary chi squares, when there are a large number of levels for the test factors, without the burden of computation that the exact method would entail. Both methods calculate the approximate summary chi square as a sum of squares of approximately orthogonal standardized deviates.

In the first method one computes an uncorrected chi square w'th 1 degree of freedom for the difference of the first level from all the remaining levels combined (the same first step as in the illustration for the three-level case). Discarding the data from the first level, a second chi square is computed for the difference between the second test-factor level and the remaining levels combined. This is done successively up to and including the last two remaining levels. The approximate summary chi square is then the sum of the separate chi squares with the number of degrees of freedom being one less than the number of test levels.

Exactly orthogonal standardized deviates would be obtained if, in the summary analysis, as each successive total deviation from expectation were evaluated, it was adjusted for its multiple regression on the preceding deviations, and then standardized by the adjusted variance. This, of course, would no longer be a simplified approximate procedure. However, it can be shown that for a single classification, in the multiple regression of any deviation from expectation on any subset of deviations, the regression coefficients will all be equal; the multiple regression on the set of deviations will be the same as the simple regression on their sum. The equality of regression coefficients, while holding true exactly for deviations in the separate subclassifications, will hold only approximately for the total deviations from expectation (it would hold exactly if equal numbers of individuals were observed from level to level at each subclassification). Nevertheless, this result suggests that approximately orthogonal deviates would be obtained if, in evaluating each successive total deviation, it were adjusted for the cumulative total of deviations already evaluated. Computing procedures to accomplish this can readily be devised.

Both approximate chi-square procedures just outlined, which may have merit when more than three groups are being compared simultaneously, should, in theory, yield linear combinations of independent chi squares. While testing the chi-square values obtained as though they were exact is not likely to be too inappropriate, it may be more correct to obtain a modified number of degrees of freedom, along the lines suggested by Satterthwaite (*47*) for problems involving such linear combinations. What the modified number of degrees of freedom would be has not been investigated by us, and it may prove as easy to apply the exact chi-square procedure, indicated later, as to determine the appropriate degrees of freedom for the approximate chi square.

It is of interest that a somewhat similar task of obtaining an appropriate summary chi square appears in the birth-order problems described by Halperin (*48*). There, it was necessary to compare a set of total observations (across family sizes) with a set of total expectations, one for each birth order. Halperin described a matrix-inversion procedure for reducing the set of correlated deviations into a summary chi square. In that problem it can be shown that all the regression coefficients are equal in the multiple regression of the deviation at a particular birth order on the set of deviations at all succeeding birth orders. The second approximate method described previously for the present problem could thus be used exactly for the birth-order problem, permitting simplified computation of chi square. The procedure indicated by Halperin has the advantage of generality and could be applied to the current and related problems, if one obtained all the necessary variances and covariances and inverted the resulting matrix.

References

(*1*) Snow, J.: On the mode of communication of cholera. *In* Snow on Cholera. New York, The Commonwealth Fund, 1936, pp. 1–139.

(*2*) Holmes, O. W.: The contagiousness of puerperal fever. *In* Medical Classics. Baltimore, Williams & Wilkins Co., vol. 1, 1936, pp. 211–243.

(*3*) Stern, R.: Nota sulle ricerche del dottore Tanchon intorno la frequenza del cancro. Annali Universali di Medicina 110: 484–503, 1844.

(*4*) Stocks, P., and Campbell, J. M.: Lung cancer death rates among non-smokers and pipe and cigarette smokers. Brit. M. J. 2: 923–929, 1955.

(*5*) Wynder, E. L., and Cornfield, J.: Cancer of the lung in physicians. New England J. Med. 248: 441–444, 1953.

(*6*) Lane-Claypon, J. E.: A further report on cancer of the breast, with special reference to its associated antecedent conditions. Rept. Publ. Health & M. Subj., No. 32, 1926, pp. 1–189.

(*7*) Clemmesen, J., Lockwood, K., and Nielsen, A.: Smoking habits of patients with papilloma of urinary bladder. Danish M. Bull. 5: 123–128, 1958.

(*8*) Denoix, P. R., and Schwartz, D.: Tobacco and cancer of the bladder. (Bulletin de L'Association francaise pour l'étude du Cancer.) Cancer 43: 387–393, 1956.

(9) Lilienfeld, A. M., Levin, M. L., and Moore, G. E.: The association of smoking with cancer of the urinary bladder in humans. A.M.A. Arch. Int. Med., 1956.

(10) Mustacchi, P., and Shimkin, M. B.: Cancer of the bladder and infestation with *Schistosoma hematobium.* J. Nat. Cancer Inst. 20: 825–842, 1958.

(11) Lilienfeld, A. M.: The relationship of cancer of the female breast to artificial menopause and marital status. Cancer 9: 927–934, 1956.

(12) Lilienfeld, A. M., and Levin, M. L.: Some factors involved in the incidence of breast cancer. *In* Proc. Third National Cancer Conference. Philadelphia, J. B. Lippincott Co., 1957, pp. 105–112.

(13) Segi, M., Fukushima, I., Fujisaku, S., Kurihara, M., Saito S., Asano, K., and Kamoi, M.: An epidemiological study on cancer in Japan. Gann Supp. 48, 1957.

(14) Dunham, L. J., Thomas, L. B., Edgcomb, J. H., and Stewart, H. L.: Some environmental factors and the development of uterine cancers in Israel and New York City. To be published in Acta Unio internat. contra cancrum.

(15) Stocks, P.: Cancer of the uterine cervix and social conditions. Brit. J. Cancer 9: 487–494, 1955.

(16) Wynder, E. L., Cornfield, J., Schroff, P. D., and Doraiswami, K. R.: A study of environmental factors in carcinoma of the cervix. Am. J. Obst. & Gynec. 68: 1016–1052, 1954.

(17) Mills, C. A., and Porter, M. M.: Tobacco smoking habits and cancer of the mouth and respiratory system. Cancer Res. 10: 539–542, 1950.

(18) Wynder, E. L., Bross, I. J., and Day, E.: A study of environmental factors in cancer of the larynx. Cancer 9: 86–110, 1956.

(19) Manning, M. D., and Carroll, B. E.: Some epidemiological aspects of leukemia in children. J. Nat. Cancer Inst. 19: 1087–1094, 1957.

(20) Breslow, L., Hoaglin, L., Rasmussen, G., and Abrams, H. K.: Occupations and cigarette smoking as factors in lung cancer. Am. J. Pub. Health 44: 171–181, 1954.

(21) Doll, R., and Hill, A. B.: A study of the aetiology of carcinoma of the lung. Brit. M. J. 2: 1271–1286, 1952.

(22) Levin, M. L.: Etiology of lung cancer; present status. New York J. Med. 54: 769–777, 1954.

(23) Sadowsky, D. A., Gilliam, A. G., and Cornfield, J.: The statistical association between smoking and carcinoma of the lung. J. Nat. Cancer Inst. 13: 1237–1258, 1953.

(24) Watson, W. L., and Conte, A. J.: Lung cancer and smoking. Am. J. Surg. 89: 447–456, 1955.

(25) Wynder, E. L., and Graham, E. A.: Tobacco smoking as possible etiologic factor in bronchiogenic carcinoma. J.A.M.A. 143: 329–336, 1950.

(26) Wynder, E. L., Bross, I. J., Cornfield, J., and O'Donnell, W. E.: Lung cancer in women. New England J. Med. 255: 1111–1121, 1956.

(27) Haenszel, W., Shimkin, M. B., and Mantel, N.: A retrospective study of lung cancer in women. J. Nat. Cancer Inst. 21: 825–842, 1958.

(28) Aird, I., Bentall, H. H., and Roberts, J. A. F.: A relationship between cancer of stomach and the ABO blood groups. Brit. M. J. 1: 799–801, 1953.

(29) Buckwalter, J. A., Wohlwend, C. B., Colter, D. C., Tidrick, R. T., and Knowler, L. A.: The association of the ABO blood groups to gastric carcinoma. Surg. Gynec. & Obst. 104: 176–179, 1957.

(30) Kraus, A. S., Levin, M. L., and Gerhardt, P. R.: A study of occupational associations with gastric cancer. Am. J. Pub. Health 47: 961–970, 1957.

(31) Cornfield, J.: A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. J. Nat. Cancer Inst. 11: 1269–1275, 1951.

(32) Dorn, H. F.: Some applications of biometry in the collection and evaluation of medical data. J. Chron. Dis. 1: 638–664, 1955.

(*33*) NEYMAN, J.: Statistics—servants of all sciences. Science 122: 3166, 1955.

(*34*) BERKSON, J.: Limitations of the application of fourfold table analysis to hospital data. Biometrics Bull. 2: 47–53, 1946.

(*35*) WHITE, C.: Sampling in medical research. Brit. M. J. 2: 1284–1288, 1953.

(*36*) GREENWOOD, M., and YULE, G. U.: On the determination of size of family and of the distribution of characters in order of birth from samples taken through members of the sibships. Roy. Stat. Soc. J. 77: 179–197, 1914.

(*37*) HAENSZEL, W.: Variation in incidence of and mortality from stomach cancer with particular reference to the United States. J. Nat. Cancer Inst. 21: 213–262, 1958.

(*38*) VIDEBAEK, A., and MOSBECH, J.: The aetiology of gastric carcinoma elucidated by a study of 302 pedigrees. Acta med. scandinav. 149: 137–159, 1954.

(*39*) WHELPTON, P. K., and FREEDMAN, R.: A study of the growth of American families. Am. J. Sociol. 61: 595–601, 1956.

(*40*) LEVIN, M. L., GOLDSTEIN, H., and GERHARDT, P. R.: Cancer and tobacco smoking. J.A.M.A. 143: 336–338, 1950.

(*41*) LEVIN, M. L., KRAUS, A. S., GOLDBERG, I. D., and GERHARDT, P. R.: Problems in the study of occupation and smoking in relation to lung cancer. Cancer 8: 932–936, 1955.

(*42*) LILIENFELD, A. M.: Possible existence of predisposing factors in the etiology of selected cancers of nonsexual sites in females. A preliminary inquiry. Cancer 9: 111–122, 1956.

(*43*) WINKELSTEIN, W., JR., STENCHEVER, M. A., and LILIENFELD, A. M.: Occurrence of pregnancy, abortion and artificial menopause among women with coronary artery disease: a preliminary study. J. Chron. Dis. 7: 273–286, 1958.

(*44*) BILLINGTON, B. P.: Gastric cancer—relationships between ABO blood-groups, site, and epidemiology. Lancet 2: 859–862, 1956.

(*45*) SCHWARTZ, D., and ANGUERA, G.: Une cause de biais dans certaines enquêtes médicales: le temps de séjour des malades a l'hôpital. Communication à l'Institut International de Statistique, 30ème Session. Stockholm, 1957.

(*46*) CORNFIELD, J.: A statistical problem arising from retrospective studies. Proc. Third Berkeley Symposium on Mathematical Statistics and Probability 4: 135–148, 1956.

(*47*) SATTERTHWAITE, F. E.: Synthesis of variance. Psychometrika 6: 309–316, 1941.

(*48*) HALPERIN, M.: The use of X^2 in testing effect of birth order. Ann. Eugenics 18: 99–106, 1953.

ff B, Gordon L, Haendel S,
:h LK, Carrano AV: The
somal constitution of human
elected for motility. Fertil
'86;46:686–690.
TA, Martin RH, Hoar DI:
ient of the sephadex tech-
or selection of X chromo-
earing human sperm by anal-
perm chromosomes, deoxy-
leic acid and Y-bodies. Fertil
989;52:829–835.
ff BF, Gordon LA, Haendel
r S, Moore DI, Gledhill BL:
omosome ratios determined
otypic analysis in albumin-
human sperm. Fertil Steril
:678–685.
U, Yanagimachi R: Sperm
some analysis as a new sys-
test human X- and Y-sperm
ion. Gamete Res 1987;17:
chi Y, Mikamo K: An im-
efficient method for analys-
ian sperm chromosomes us-
a-free hamster ova. Am J
enet 1986;38:724–740.
do C, Benet J, Genesca A,
J, Caballin MR, Miro R,
e J: Human sperm chromo-
Hum Reprod 1988;3:133–
RH. Sex ratio among sperm
m J Hum Genet 1990;47:349–

Original Paper

Hum Hered 1992;42:337–346

A Haplotype-Based 'Haplotype Relative Risk' Approach to Detecting Allelic Associations

Joseph D. Terwilliger
Jurg Ott

Department of Genetics and
Development,
Columbia University,
New York, N.Y., USA

Key Words
Linkage disequilibrium
Haplotype relative risk
Allelic association,
Chi-square tests

Abstract

A novel variation of the Haplotype Relative Risk (HRR) of Rubinstein et al. [Hum Immunol 1981;3:384] is proposed, in order to glean increased information about linkage disequilibrium or allelic associations by analyzing haplotype-based data rather than genotypic data. It is shown that statistical tests based on our design give much higher power than those based on the original HRR approach. Several additional nonparametric tests based on the same data are analyzed, and power is computed for each of them. Further, parametric likelihood methods are applied to testing linkage equilibrium, and estimating δ, the coefficient of linkage disequilibrium, from the same data.

Introduction

Allelic associations between etiologically unrelated traits were originally detected in humans through observations at the genotypic level. In the 1950s, it was noticed that in individuals with certain diseases there were significant excesses of certain blood groups. Aird et al. [1, 2] demonstrated the presence of a significant association between blood group A and stomach cancer, and between blood group O and peptic ulcer, while Pike and Dickens [3] found such an association between blood group O and toxemia of pregnancy, and McConnell et al. [4] studied associations between blood groups and carcinoma of the lung. Woolf [5] then proposed his Relative Risk statistic to compare the incidence rates in given blood groups in a case control type of study, in which one would collect a sample of people with the disease and compare the observed frequency of the 'risk allele' with its frequency in a separate sample of healthy individuals (or population frequency, if known).

One problem with this method is that there is no way of knowing whether a significant re-

Joseph D. Terwilliger
722 West 168th St. Box 58
New York, NY 10032 (USA)

sult is biologically meaningful or just a consequence of having the case and control samples taken from different genetic populations in which the frequency of the risk allele is different and therefore, no real association exists. To attempt to circumvent this problem, Rubinstein et al. [6] proposed the Haplotype Relative Risk (HRR) statistic, based on earlier work of C.A.B. Smith, to ensure that the control and disease samples were well-matched, from the same population, so that any observed association would have to be due to a real allelic association of some sort. This experimental design has also been used in the haplotype frequency difference statistic of Seuchter et al. [7].

Experimental Design

H = Marker allele with which disequilibrium is hypothesized.
$\overline{H}$ = Any allele other than H at the marker locus.
δ = Gametic linkage desiquilibrium coefficient;
= P(AB gamete) – P(A)P(B) (A at one locus B at the other).
Θ = Recombination fraction between marker and disease loci.
p = Gene frequency of the disease allele.
q = Gene frequency of the H allele.
n = Sample size.

Table 1. Data collected in a haplotype relative risk study (either HHRR, or GHRR)

Transmitted	Not transmitted		Total
	H	$\overline{H}$	
H	A	B	W
$\overline{H}$	C	D	X
Total	Y	Z	N

In the 2×2 table shown here, each cell corresponds to one parent. In the HHRR, each parent transmits one allele, and not the other, and can thus be classified by which allele was, and which was not transmitted to the affected offspring. In the GHRR, each set of parents has 4 alleles, 2 of which are transmitted to the affected child, and 2 which are not. If the child contains 1 or 2 H alleles, we say H was transmitted, and if there is an H allele in the remaining 2 alleles, we say that H was not transmitted. Thus, each family either transmits H or $\overline{H}$, and has either H or $\overline{H}$ among the nontransmitted alleles, and can therefore also be characterized by one cell of this table.

Table 2. Haplotype relative risk

	H	$\overline{H}$	Total
Transmitted	W	X	N
Not transmitted	Y	H	N
Total	W+Y	X+Z	2N

The data in this table are taken directly from the marginals of table 1, and represent the form of the originally proposed GHRR statistic. This table, of course, can be filled with either haplotype- or genotype-based data. All variable names are the same as in table 1.

In order to be sure one has matched control and disease samples, Rubinstein et al. [6] proposed using data from nuclear families with one affected offspring to test for deviations from linkage equilibrium. They recommended using the affected offspring's genotype (made up of alleles transmitted from parents to the affected child) at a marker locus as the 'case' sample, and an artificial genotype made up of the alleles not transmitted to the child from its parents as the 'control' sample in an association test. Then they used such data to test whether the H allele was present equally frequently in diseased individuals' genotypes, and the nontransmitted control genotypes. For example, in a family with unaffected parents with genotypes G/H and I/J at the marker locus, and an affected child with marker genotype H/I, the transmitted genotype would be H/I, and the artificial nontransmitted genotype would be G/J. Since they were only interested in whether H was present or absent from the genotypes, in this example we have H transmitted, and $\overline{H}$ not transmitted (genotype G/J does not contain H). For every such nuclear family there would be one such observation. One can then tabulate such observations in the form of table 1. The example family above would fall in cell B. Ott [8] demonstrated that under the null hypothesis of δ = 0, the transmitted and nontransmitted alleles are independently associated, and thus we can treat our transmitted and nontransmitted samples independently and represent them in the form of table 2 (marginals of table 1). Then a standard χ^2 test of independence on this table can be shown to be a valid χ^2 test of the hypothesis $\delta = 0$. This is the test proposed by Rubinstein et al. [6] to guarantee the control and disease samples are genetically well-matched.

As is shown below, the statistical method of Rubinstein et al. [6] does not take advantage of all the information present in the data. Their method lumps H/H homozygotes and H/$\overline{H}$ heterozygotes together as H genotypes. However, since under the null hypothesis the two parental genotypes are independent, it is possible to treat each parent as an independent observation, and merely look at the fate of each parental marker allele. So, in the example family above, there would be one observation of H transmitted, G not transmitted, and one observation of I transmitted, J not transmitted, which in table 1 (now referring to alleles, not genotypes), would contribute one observation to cell B, and one observation to cell D. Again, for theoretical reasons given by Ott [8], transmitted and nontransmitted alleles are independent for each other, and can be collapsed, as in the Rubinstein case, into table 2, in which the example family would contribute one observation to cell W, one to cell X, and two to cell Z, the marginal values of table 1. We are thus using more of the information present in the family, obtaining twice as many observations from the same amount of data.

Recessive Disease

Haplotype-Based versus Genotype-Based HRR χ^2 Tests

We first compared the power of our haplotype-based HRR (HHRR) statistic with the genotype-based HRR (GHRR) of Rubinstein et al. [6]. The test we applied to each data set is essentially a χ^2 test of independence on table 2 for the haplotype-based data (HHRR test), and for the equivalent genotype-based table (GHRR test) in which discrimination is between genotypes with no H allele, and those with at least one (possibly two). Power calculations were performed for each test, assuming a recessive disease with no phenocopies (penetrance is irrelevant to the calculations, according to Ott [8]), by analytically computing the probability of a significant χ^2 test result ($\chi^2 > 3.84$ at the 0.05 level) for different combinations of δ/p (δ and p are completely confounded according to Ott [8]), q, and Θ. Power curves for these two tests (n = 100 families, q = 0.5) are given in figure 1 for varying true values of Θ and δ/p. In all the numerical cases we considered, the HHRR test was more powerful than the GHRR approach of Rubinstein et al. [6]. This is intuitively satisfying, since the HHRR approach discriminates between H/H homozygotes and H/$\overline{H}$ heterozygotes, while the GHRR does not. Thus, our approach uses all of the information in the data, where the traditional GHRR does not.

The test of independence on table 2 is a test of E[W] = E[Y]. However, W and Y are obtained from the marginals of table 1. So, when we are testing E[W] = E[Y], we are essentially testing E[A + B] = E[A + C], which is the same as E[B] = E[C]. Clearly this is expected under the null hypothesis of no disequilibrium. Using the data from table 1, the HHRR χ^2 is computed as

$$\frac{2N(B-C)^2}{(2A+B+C)(N-2A-B-C)} = \frac{2N(WZ-XY)^2}{(W+X)(W+Y)(X+Z)(Y+Z)},$$

the standard χ^2 test of independence on a 2 × 2 table. This is a valid χ^2 test, of the form $(B-C)^2/\mathrm{Var}[B-C]$, since $\mathrm{Var}[B-C] = 2Nq(1-q)$, which is estimated by $2N[(2A+B+C)/(2N)][1-(2A+B+C)/(2N)]$. The power is shown graphically in figure 2 for n = 50 families (for comparison with other haplotype-based tests below).

McNemar Tests

Since our null hypothesis is B = C in a paired sampling (transmitted allele, nontransmitted allele) test, one's first intuition might

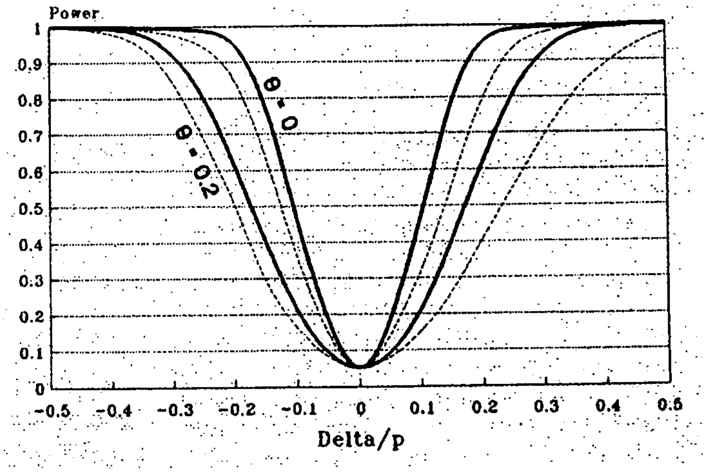


Fig. 1. Power curves (analytically computed) for χ^2 tests based on the haplotype- (——) and genotype-based (----) HRR designs (100 families), for q = 0.5. If p = 0.5, then all values of δ/p shown are possible. For other values of p, different restrictions apply, but have no effect on the power curve. The upper two lines are for the power of the test when Θ = 0, and the lower set of two lines correspond to Θ = 0.20. Note that the haplotype-based design yields higher power for all true values of Θ and δ/p.

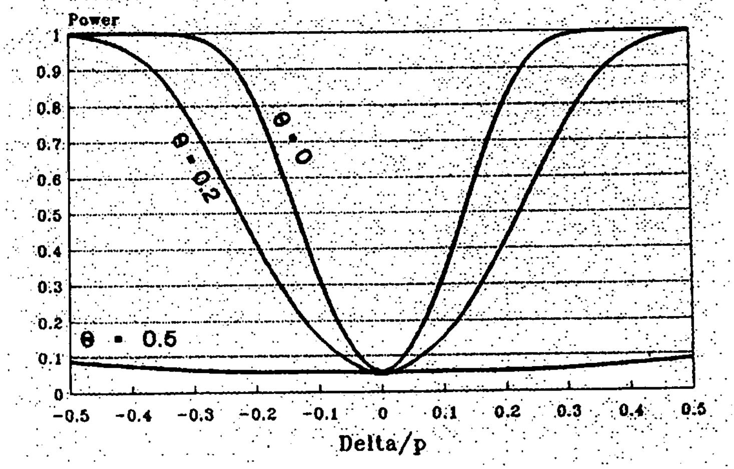
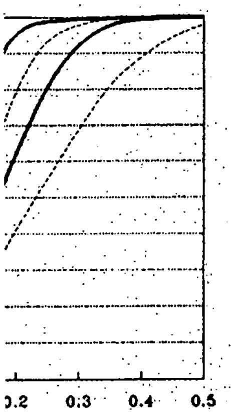


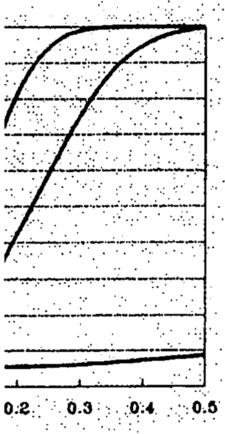
Fig. 2. Power curves (analytically computed) for the HHRR test (50 families) for q = 0.5, with Θ = 0 (upper curve), 0.2 (middle curve) and 0.5 (lower curve).

be to apply a McNemar test, $(B-C)^2/(B+C)$. In order for this to also be a valid χ^2 test, $(B+C)$ would have to be an estimate of the variance of $(B-C)$, which we already have shown to be $2Nq(1-q)$. Our HHRR χ^2 test uses all the data to estimate q, including the information from homozygous individuals, while in the McNemar test, all homozygotes are ignored, and the variance is estimated as $(B+C)$. Clearly $E[C] = E[B] = Nq(1-q)$ under the null hypothesis $(\delta = 0)$, so $(B+C)$ then estimates $2Nq(1-q)$. However, in every numerical case we considered, this test was less powerful than the HHRR test, as shown in figure 3, due to the fact that the HHRR uses all of the data to estimate the variance, while the McNemar uses only the information from heterozygous parents.

Independence Tests

An interesting result of Ott [8] is that transmitted and nontransmitted alleles are inde-





ʼer, in every numer-
ıis test was less pow-
t, as shown in figure
: HHRR uses all of
variance, while the
formation from het-

: Ott [8] is that trans-
ed alleles are inde-

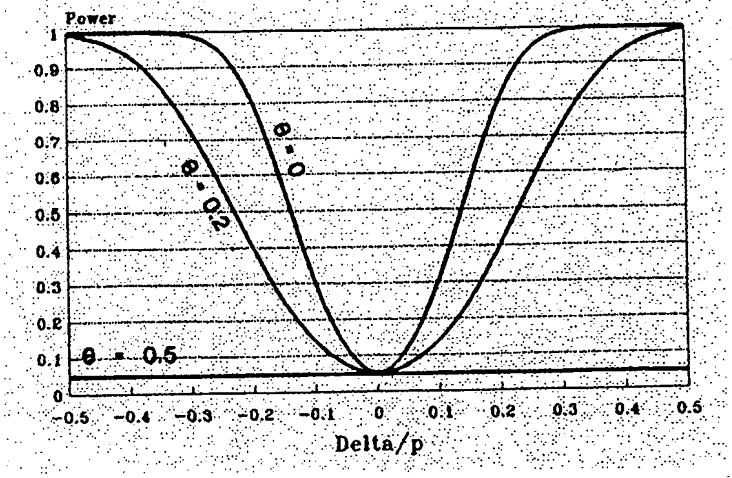


Fig. 3. Power curves (analytically computed) for the haplotype-based McNemar (HMCN) test (50 families) for q = 0.5, with Θ = 0 (upper curve), 0.2 (middle curve), and 0.5 (lower curve).

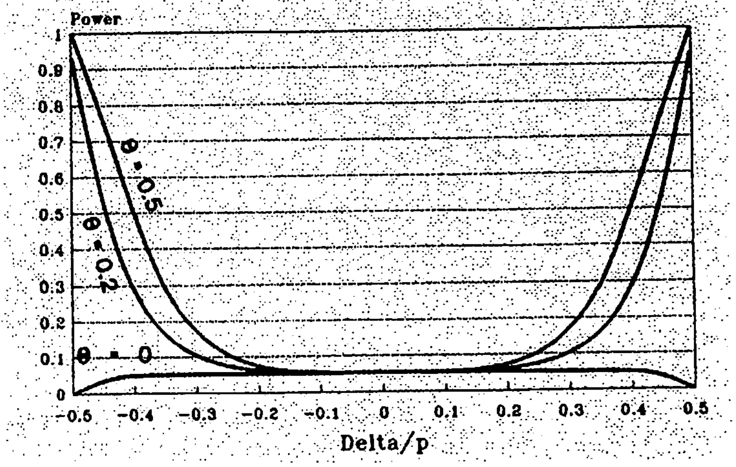


Fig. 4. Power curves (analytically computed) for the haplotype-based independence test (HIND) for 50 families, q = 0.5, and Θ = 0 (lower curve), 0.2 (middle curve), and 0.5 (upper curve).

pendent when $\delta = 0$ or when $\Theta = 0$. In light of this, one could use an independence test on table 1 as a test of $\delta = 0$, though clearly when Θ is close to 0, this test should not be useful. This test is just that (AD–BC) = 0. Therefore, the test should be $(AD–BC)^2/Var(AD–BC)$, which is the standard χ^2 test of independence on a 2 × 2 table, $N(AD–BC)^2/(WXYZ)$. Power was analytically computed for this test, under the recessive model, for various true values of q, δ/p, and Θ, which are graphically presented in figure 4. In this test, the power increases as Θ increases, just the opposite behavior from the HHRR and McNemar tests. This test may thus be a useful way to use such nuclear family data to test $\delta = 0$ when Θ is known to be quite large, since when $\Theta = 0.5$, the HHRR tends to 0 [8].

This independence test, however, fails to impose the restriction that the frequency of the H allele be equal in both the transmitted and nontransmitted samples. To include this

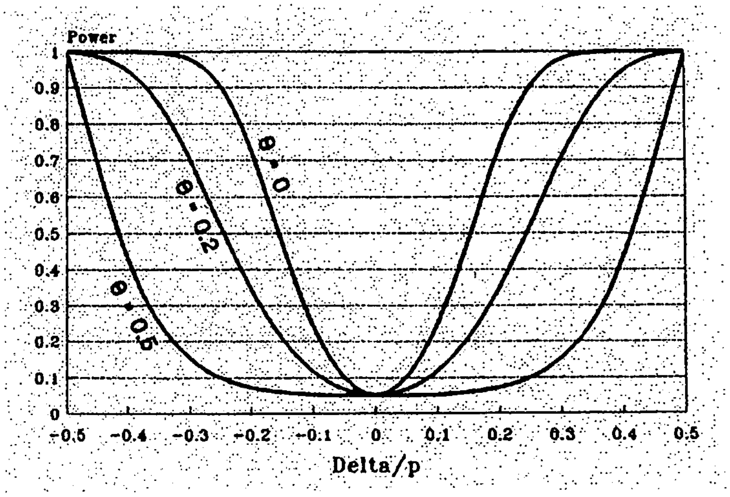


Fig. 5. Power curves (analytically computed) for the test of fit to the expected multinominal proportions (HIID) of haplotype-based data for 50 families, q = 0.5, and Θ = 0 (upper curve), 0.2 (middle curve), and 0.5 (lower curve).

information, one could test the fit of the counts of A, B, C, and D to their expected multinominal proportions (each observation is clearly independent) as follows: $\Sigma(O-E)^2/E$, which is equal to

$$\frac{(A-N\hat{q}^2)^2}{N\hat{q}^2}+\frac{[B-N\hat{q}(1-\hat{q})]^2}{N\hat{q}(1-\hat{q})}+\frac{[C-N\hat{q}(1-\hat{q})]^2}{N\hat{q}(1-\hat{q})}$$

$$+\frac{[D-N(1-\hat{q})^2]^2}{N(1-\hat{q})^2}, \text{ where } \hat{q}=\frac{2A+B+C}{2N}.$$

This test follows a χ^2 distribution with 2 df, since we had 4 cell counts, but fixed the sum $A+B+C+D=N$, and estimated q from the data. This test is very powerful over a large range of values of δ/p, q, and Θ, as shown in figure 5, and thus provides a useful general test for disequilibrium.

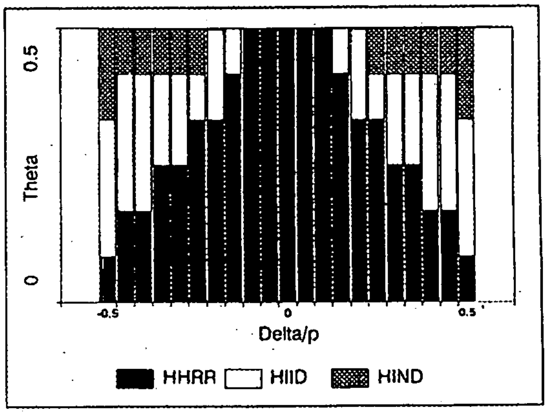
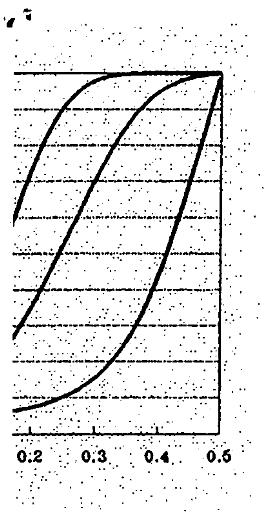
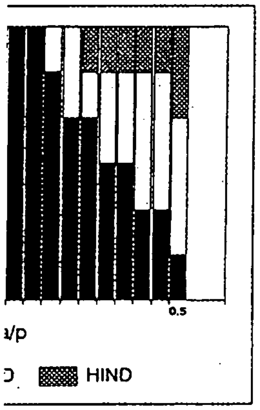


Fig. 6. Graph showing, for all possible values of Θ and δ/p, and fixed q = 0.5, which among three tests is the most powerful (50 families). The values of the power are not shown, but are given in fig. 2–5 (HMCN is never the most powerful).

Relative Power of Nonparametric Approaches

Each of the tests described above has different properties which make it useful. However, the question remains as to which test should be used in a given situation. To answer that question, for each combination of Θ, δ/p, and q, we determined which test gave maximal power for a sample size of 50 families. The results are presented graphically in figure 6. In this figure, for fixed q, we considered all possible combinations of δ/p and Θ, and determined which test gave maximal power (analytically computed). Then for each point (δ/p, Θ) the most powerful test is indicated. To see ex-





r all possible values of Θ
hich among three tests is
lies). The values of the
given in fig. 2–5 (HMCN

ich test gave maximal
›f 50 families. The re-
hically in figure 6. In
e considered all pos-
'p and Θ, and deter-
ıximal power (analyt-
›r each point (δ/p, Θ)
indicated. To see ex-

actly what the power was, the reader is referred to the power curves already presented for each test. Some interesting patterns can be seen in this figure, but it should be used only in conjunction with the actual values of the power shown in figures 2–5, for often the difference is small between tests. However, over the most relevant ranges of δ/p and Θ, for all q, the HHRR test is the most powerful. In light of this, and the relative implausibility of strong disequilibrium when Θ is large, the HHRR test should be the general nonparametric test of choice, both for its power, and its simplicity.

Parametric Likelihood Ratio Tests

If one knows the model of the disease, one could do a parametric likelihood ratio test analysis, based on theoretical probabilities of each type of parent under a fixed model. Table 2 of Ott [8] provides such parametric values for the case of a recessive disease. The difficulty here is three fold. First, one needs to have an accurate parametric model for the disease, and compute the parametric probabilities of each cell of table 1. This process is very tedious (except for the recessive model described by Ott [8]), and depends heavily on the disease model. Secondly, one needs to maximize the likelihood of the data over all the parameters, Θ, (δ/p), and q, and then again maximize the likelihood, fixing δ = 0. This would give us the following likelihood ratio: $L(\hat{\delta}/p, \hat{\Theta}, \hat{q})/L(\delta/p = 0, \hat{\Theta}, \hat{q})$. Normally, one can treat $2 \times \ln(LR)$ as a χ^2 random variable, with the number of degrees of freedom being the difference in free parameters in numerator and denominator of the likelihood ratio, which would appear to be 1 in this case. However, when δ = 0, Θ disappears as a parameter, as shown by Ott [8]. When a parameter disappears under the null hypothesis, it is a degenerate situation, and so the statistic does not satisfy the criteria for χ^2. As the distribution is unclear, this test becomes very awkward to interpret, and presents a situation analogous to the degenerate likelihood ratio test for linkage in the presence of heterogeneity [9]. For this reason, combined with the enormous computer time involved, power was not calculated for this approach.

For general pedigree data (including nuclear families with multiple offspring), with a fixed-disease model, parametric likelihood ratio tests are tractable using any linkage analysis program, like ILINK of the LINKAGE package. One need only maximize the likelihood over Θ, q, p, and δ for the numerator, and again maximize the likelihood for the denominator over Θ, q, and p, fixing δ = 0. This would then be a valid, and powerful general likelihood ratio test of δ = 0, $2 \times \ln[L(\hat{\Theta}, \hat{\delta}, \hat{p}, \hat{q})/L(\hat{\Theta}, \delta = 0, \hat{p}, \hat{q})]$. It is important to remember that when using this method, the maximum likelihood estimates of the haplotype frequencies will reflect the sample frequency of the disease allele, which is not an accurate reflection of its population frequency. One must be sure to weight disease and control haplotypes. accordingly. For example, if our haplotype frequency estimates are $\hat{P}(Hd)$, $\hat{P}(\overline{H}d)$, $\hat{P}(HD)$, $\hat{P}(\overline{H}D)$, and we know the true gene frequency of the d allele, p_d, we can compute adjusted haplotype frequency estimates as

$$\hat{P}(Hd) = \left(\frac{\hat{P}(Hd)}{\hat{P}(Hd)+\hat{P}(\overline{H}d)}\right)(p_d),$$

and so on. Similarly, if one wanted to estimate the coefficient of disequilibrium from such ILINK estimates, it would be necessary to use the adjusted estimates described above, yielding an adjusted estimate of

$$\hat{\delta} = (\hat{\delta})\frac{p_d(1-p_d)}{\hat{p}_d(1-\hat{p}_d)},$$

where $\hat{\delta} = \hat{P}(Hd)\,\hat{P}(\overline{H}D) - \hat{P}(HD)\,\hat{P}(\overline{H}d)$, and $\hat{p}_d = \hat{P}(Hd) + \hat{P}(\overline{H}d)$.

An ad hoc method sometimes used in general pedigrees is to assume the absence of recombination, and determine the haplotypes of each founder, between marker and disease, as a way to insure the control (nondisease) haplotype are from the same genetic population as the disease haplotypes. This ad hoc approach has been applied, for example, in cystic fibrosis [10]. It assumes an absence of recombination, and its statistical properties are, in general, unclear, especially in cases where Θ is actually greater than zero. Another problem is that it is not always possible to uniquely and accurately determine all founder haplotypes. Censoring such indiscernible cases in some instances can be shown to lead to a statistical bias. In light of all of this, if one wants to use general pedigree data to test and quantify disequilibrium, the likelihood ratio test with ILINK described above is the test of choice, as it is more general and powerful, and has well-characterized statistical properties.

Nonrecessive Case

All of our results above were obtained for the case of a recessive disease. However, when other more complicated models prevail, the situation becomes unclear. While under any model we choose for the disease, the above tests are valid tests of $\delta = 0$ (since this implies no association between the disease and the marker locus), the effect on the power of our testing procedures is not so clear. When dealing with a recessive disease, a lot of additional information about linkage disequilibrium is obtained by looking at each parent separately, since each parent transmits a disease allele to the affected offspring, but the situation is less clear when there is a different model. For a dominant disease, with one affected parent, and one affected child, one can just consider the affected parent, and his or her transmitted and nontransmitted alleles, and base a test on the same procedure as above. The effect would be that there would be only one observation per family instead of two in the recessive case (where we know the parents to be heterozygous for the disease), and there is possible noise when the unaffected parent actually transmits the disease to the offspring, though this should be very rare.

In the case of dominant reduced-penetrance disease, in which neither parent is affected, clearly at least one parent must carry the disease-predisposing allele, though we cannot discern which one. In this situation, one parent will transmit the disease allele (in putative disequilibrium with the marker), and the other parent will transmit the normal allele. This adds noise to our system. One would expect the Rubinstein method to be less sensitive to this noise, since it doesn't distinguish between heterozygotes and homozygotes for the H allele.

Power calculations were approximated for this situation by simulation. A simplified model was considered in which one parent was forced to transmit the disease allele to the affected child, while the other parent was assumed to be homozygous unaffected (a reasonable assumption for small p). In this case, δ and p are no longer completely confounded, so we had to treat p, q, δ, and Θ as separate parameters. Then, 20,000 sets of 100 such nuclear families with 2 unaffected parents and one affected offspring were simulated under various assumptions on p, q, δ, and Θ. For each set of 100 families, the HHRR and GHRR were calculated. Then the number of significant results for each test at the 0.05 level ($\chi^2 \geq 3.84$) was counted to estimate the power of each test, which is graphed in figure 7. An interesting situation arises here, where the HHRR is much more powerful for negative values of δ, but for positive values of δ they are just about equal in power, with the GHRR being slightly

:s, and base a test on
above. The effect
d be only one obser-
of two in the reces-
w the parents to be
ıease), and there is
naffected parent ac-
ıse to the offspring,
y rare.
nant reduced-pene-
neither parent is af-
.e parent must carry
allele, though we
.e. In this situation,
:he disease allele (in
ith the marker), and
ısmit the normal al-
r system. One would
thod to be less sensi-
: doesn't distinguish
nd homozygotes for

re approximated for
ıtion. A simplified
/hich one parent was
ease allele to the af-
ther parent was as-
; unaffected (a rea-
ıall p). In this case, δ
pletely confounded,
nd Θ as separate pa-
s of 100 such nuclear
parents and one af-
ılated under various
d Θ. For each set of
nd GHRR were cal-
:r of significant re-
).05 level ($\chi^2 \geq 3.84$)
the power of each
igure 7. An interest-
vhere the HHRR is
ıegative values of δ,
; they are just about
;HRR being slightly

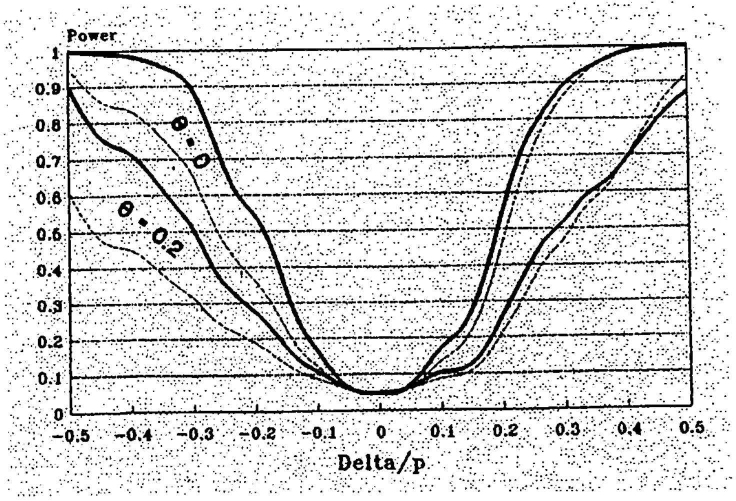


Fig. 7. Power curves (simulated) for the HHRR (—) and GHRR (----) tests with a dominant disease (reduced penetrance) and two unaffected parents, for q = 0.5, p = 0.01, and 100 families, based on 20,000 replicates. The upper curves represent Θ = 0, and the lower curves Θ = 0.2. In most cases, the HHRR is shown to be much more powerful than the GHRR.

more powerful for very extreme values of δ. The HHRR test is also more powerful than the other haplotype-based nonparametric tests over most of the reasonable sample space. The HHRR is more powerful than the GHRR in all recessive situations, dominant situations with δ < 0, and about equally powerful with the GHRR in dominant situations with extremely positive δ. Further, the HHRR can take advantage of dominant situations with one affected parent, while the GHRR cannot. Therefore, we recommend using the HHRR as the nonparametric test of choice in general.

Discussion

When doing an association study, it is often difficult to find genetically well-matched cases and control samples. The HRR approach of using transmitted and nontransmitted alleles from the same parent as case and control samples ensures that they are genetically well-matched [11]. Further, the case and control samples are shown to be independent under the null hypothesis of δ = 0. In light of this, HRR-type methods should be increasingly more important as geneticists try to map complex diseases, by looking for associations with candidate genes for example. In such a case, if the candidate gene is correct, Θ would be equal to 0, and these methods would achieve maximal power to detect the associations. Further, the built-in genetic control should provide a solution to the often difficult task of finding a valid control sample, and should allow people to have more faith in the validity of such association studies.

The approach presented here extracts further information about disequilibrium from the data used in the original GHRR approach, and thus presents a more powerful way to detect such associations in the absence of a parametric model. Given a parametric model, two likelihood-based methods were discussed as well. However, from the results of our power calculations, our HHRR seems to be the best general nonparametric test considered for detecting such associations with this experimental design over the most biologically plausible ranges of δ and Θ.

Acknowledgements

This work was supported by the National Institute of Mental Health (grant MH44292), the MacArthur Foundation Network I (Task Force on the Development of Analytic Strategies for Linkage Studies of Psychiatric Disorders), and the W.M. Keck Foundation. Further, the authors would like to acknowledge the contributions of Dr. Warren Ewens, whose comments are greatly appreciated.

References

1 Aird I, Bentall HH, Roberts JAF: The relationship between cancer of the stomach and the ABO blood groups. BMJ 1953;1:799–801.

2 Aird I, Bentall HH, Mehigan JA, Roberts JAF: The blood groups in relation to peptic ulceration and carcinoma of colon, rectum, breast, and bronchus. BMJ 1954;2:315–321.

3 Pike LA, Dickens AM: ABO blood groups and toxaemia of pregnancy. BMJ 1954;2:321–323.

4 McConnell RB, Clarke CA, Downton F: Blood groups in carcinoma of the lung. BMJ 1954;2:323–325.

5 Woolf B: On estimating the relation between blood and disease. Ann Hum Genet 1955;19:251–253.

6 Rubinstein P, Walker M, Carpenter C, Carrier C, Krassner J, Falk CT, Ginsburg F: Genetics of HLA disease associations. The use of the Haplotype Relative Risk (HRR) and the 'Haplo-Delta' (Dh) estimates in juvenile diabetes from three racial groups. Hum Immunol 1981;3:384.

7 Seuchter SA, Knapp M, Baur MP: Analysis of association in nuclear families; in Lynch HT, Tautu P (eds): Recent Progress in the Genetic Epidemiology of Cancer. Berlin, Springer, 1990, pp 89–94.

8 Ott J: Statistical properties of the Haplotype Relative Risk. Genet Epidemiol 1989;6:127–130.

9 Ott J: Analysis of Human Genetic Linkage. Baltimore, Johns Hopkins University Press, 1991.

10 Estivill X, Farrall M, Williamson R, Ferrari M, Seia M, Giunta AM, Novelli G, Potenza L, Dallapicolla B, Borgo G, Gasparini P, Pignatti PF, De Benedetti L, Vitale E, Devoto M, Romeo G: Linkage disequilibrium between cystic fibrosis and linked DNA polymorphisms in Italian families: A collaborative study. Am J Hum Genet 1988;43:23–28.

11 Falk CT, Rubinstein P: Haplotype Relative Risks: An easy way to construct a control sample for risk calculations. Ann Hum Genet 1987;51:227–233.

Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations

C. T. FALK AND P. RUBINSTEIN

The Lindsley F. Kimball Research Institute of The New York Blood Center, 310 E. 67th St., New York, NY 10021

SUMMARY

An alternative to Woolf's (1955) relative risk (RR) statistic is proposed for use in calculating the risk of disease in the presence of particular antigens or phenotypes. This alternative uses, as the control sample, the parental antigens or haplotypes not present in the affected child. The formulation of a haplotype relative risk (HRR) thus eliminates the problems of sampling from the same homogeneous population to form both the disease sample and an appropriate control.

We show that, in families selected through a single affected individual, where transmission of the four parental haplotypes can be followed unambiguously, the mathematical expectation of the HRR is identical to that of the RR. Since the sample formed from the 'non-affected' parental haplotypes is clearly from the same population as the disease sample, the HRR thus provides a reliable alternative to the RR. A further advantage obtains when family data are being collected as part of a study since the control sample is then automatically contained in the family material.

Data from studies of patients with insulin dependent diabetes mellitus (IDDM) are used to obtain an estimate of the risk to those with HLA antigens or phenotypes associated with IDDM using the HRR statistic. A comparison of the HRR's and RR's for these data is also presented.

INTRODUCTION

Relative risks have been used for some time to estimate the increased risk of contracting a disease, given that a certain condition (or trait) is present, over that of the group lacking the condition. This formal definition of a relative risk requires prospective information that is not easily obtained and the relative risk is often approximated by the more easily obtained cross product

$$\frac{\Pr(Q\,|\,\text{aff})\Pr(q\,|\,\text{control})}{\Pr(q\,|\,\text{aff})\Pr(Q\,|\,\text{control})},$$

where Q stands for the presence of the condition or trait and q for the lack of the condition, and the four terms are conditional probabilities as indicated. When the overall frequency of the disease in a population is low, this estimate will closely approximate the true relative risk. This odds ratio was proposed by Woolf (1955) to estimate the risk of contracting either peptic ulcers or stomach cancer for individuals of particular ABO phenotypes. Since then it has been used to calculate risks for genetic markers associated with many diseases and its most notable use has been in studying several HLA-associated diseases such as insulin dependent diabetes mellitus (IDDM), coeliac disease, multiple sclerosis and ankylosing spondylitis. Several assumptions are generally made about the underlying population from which both the disease sample and the control sample are obtained, most importantly that both samples are drawn from the same genetically homogeneous population in an unbiased way. By this we mean that the disease sample should be selected on a clear-cut ascertainment criterion, e.g. randomly chosen affected individuals with no bias pertaining to other factors, and the control sample should be a strictly random sample from the same genetic population. In practice, this latter criterion is rather difficult to fulfil and most often the control is created from conveniently available data drawn from a population thought to be somewhat closely related to that from which the disease sample was drawn.

Several years ago we proposed (Rubinstein *et al.* 1981) an alternative method for obtaining the control sample for relative risk (RR) estimations that eliminated the problems of sampling from a single homogeneous population. This method used, as a control, those parental haplotypes not present in the affected child and was therefore called the haplotype relative risk (HRR). This method has several appealing features including freedom from collection of proper control samples. Additionally, where families are to be studied anyway, collection of the family data automatically includes collection of the necessary control sample. It is, however, necessary to demonstrate that the HRR estimate has the appropriate characteristics. In this paper we will show that, assuming the 'ideal' conditions inherent in the definition of RR, namely, control and disease samples both randomly chosen from the same homogeneous random mating population, the expected value of the HRR is identical to that of the conventional RR. We will then illustrate its use in the estimation of risks for HLA antigens and phenotypes associated with IDDM.

THE MODEL

Consider a set of families that has been ascertained through a single affected child, where the relevant disease locus is closely linked to a normal polymorphic genetic marker (e.g. HLA) and where certain alleles (antigens) are associated with the disease. For purposes of concreteness, we will assume that the disease is recessively inherited, although the same arguments hold for dominance and for other inheritance models as well. Assume that the HLA haplotypes present in the parents can be followed unambiguously in transmission to the offspring and designate the two inherited by the affected child as '*a*' (paternal) and '*c*' (maternal). Thus haplotypes *a* and *c* are assumed to carry the disease allele, say '*n*'. In the special case where the child as well as both parents are *ac*, it is not certain whether the child gets the *a* from the mother or the father. However, it is still known that one *a* and one *c* haplotype were transmitted to the affected child, and thus carry the *n* allele, and that the haplotypes not passed on to the affected child were also *a* and *c*. The latter can therefore be included in the 'random sample' as described below. Now if we have truly obtained our sample as a random, singly selected sample, the two parental haplotypes not transmitted to the affected child (say *b* and *d*) will represent a random sample of haplotypes from the population at large and will thus carry the disease allele (*n*) or the normal allele (*N*) with probabilities equal to the allele frequencies in the population (say p_1 and p_2, respectively, $p_1+p_2 = 1$). The validity of this observation requires compliance with certain other assumptions including (1) that the parents are not inbred, (2) that there is no correlation within or between parental phenotypes and (3) that there is no differential fertility of the disease phenotypes.

Now assume that an antigen '*Q*' at the HLA locus is in positive linkage disequilibrium with *n*, the disease allele. We wish to calculate the relative risk to carriers of *Q* of contracting the disease. We will use as our control population the set of '*b*' and '*d*' haplotypes from our sample of disease families (that is, those haplotypes within a family not carried by the single affected proband). Using this control we will then calculate the conventional cross product odds ratio given above to obtain the haplotype relative risk (HRR). Define the relevant population frequencies as follows:

$$
\begin{aligned}
f(Q) &= q_1,\\
f(q) &= q_2 = 1-q_1 \quad \text{(where } q \text{ represents all other alleles)},\\
f(n) &= p_1,\\
f(N) &= p_2 = 1-p_1,\\
f(Qn) &= x_1 = p_1 q_1+\delta,\\
f(QN) &= x_2 = p_2 q_1-\delta,\\
f(qn) &= x_3 = p_1 q_2-\delta,\\
f(qN) &= x_4 = p_2 q_2+\delta,
\end{aligned}
$$

where δ is the measure of disequilibrium between n and Q.

We now need the four conditional probabilities necessary for the odds ratio. For the affected sample these are the same, regardless of how we choose our control.

$$
\begin{aligned}
\Pr(Q|\text{aff}) &= \frac{x_1^2+2x_1 x_3}{(x_1+x_3)^2},\\
&= \frac{p_1^2-x_3^2}{p_1^2},\\
\Pr(\text{not } Q|\text{aff}) &= \frac{x_3^2}{(x_1+x_3)^2} = \frac{x_3^2}{p_1^2}.
\end{aligned}
$$

Now since the control haplotypes will be a random sample from the population, the conditional probabilities will be:

$$
\begin{aligned}
\Pr(Q|\text{control}) &= 1-(x_3+x_4)^2 = 1-q_2^2\\
\Pr(\text{not } Q|\text{control}) &= (x_3+x_4)^2 = q_2^2.
\end{aligned}
$$

Thus the estimate of the HRR is:

$$
\begin{aligned}
\text{HRR} &= \frac{\Pr(Q|\text{aff})\Pr(\text{not } Q|\text{control})}{\Pr(\text{not } Q|\text{aff})\Pr(Q|\text{control})}\\
&= \frac{(p_1^2-x_3^2)\,q_2^2}{x_3^2(1-q_2^2)},
\end{aligned}
$$

which is identical to the equivalent expression for the conventional RR.

EXAMPLE

Using data collected for the 9th HLA Workshop (Bertrams & Baur, 1984) we looked at the sample of families, submitted for study, where a single child was affected with IDDM and where the ethnic background was caucasoid (Western European or North American). The patients

Table 1. DR *phenotypes of* IDDM *disease sample, simplex cases*

DR type	No. obs.	No. exp.
DR3, 3	6	7·8
DR3, 4	25	18·2
DR4, 4	4	10·7
DR3, X	16	19·1
DR4, X	29	22·4
DRX, X	10	11·7
Total	90	89·9

$p(DR3) = 0{\cdot}294$: $p(DR4) = 0{\cdot}344$: $p(DRX) = 0{\cdot}361$; $\alpha/\beta = (0{\cdot}278)/(0{\cdot}202) = 1{\cdot}38$.

Table 2. DR *phenotypes of 'control' sample consisting of non-affected parental haplotypes*

DR type	No. obs.	No. exp.
DR3, 3	0	0·77
DR3, 4	2	1·23
DR4, 4	0	0·49
DR3, X	13	12·26
DR4, X	10	9·76
DRX, X	48	48·49
Total	73	73·00

$p(DR3) = 0{\cdot}103$: $p(DR4) = 0{\cdot}082$; $P(DRX) = 0{\cdot}815$; $\chi^2 = 1{\cdot}79$, 2. d.f.

were categorized with respect to their HLA DR phenotypes using three distinct allelic groups *DR3*, *DR4*, and *DRX*, where *DRX* represents all other *DR* antigens except *DR3* and *DR4*. The results are shown in Table 1 with estimated allele frequencies and observed and 'Hardy–Weinberg expected' numbers for each phenotypic class. The α/β ratio of Falk *et al.* (1983) was also calculated and found to be 1·38. This ratio relates the observed frequency (α) of, say the DR3,4 phenotype, to the Hardy–Weinberg expected frequency [$\beta = 2p(DR3)p(DR4)$] in a sample of diseased individuals (Table 1). A value in excess of 1·0 is an indication that the associated susceptibility locus does not show a simple dominant or recessive mode of inheritance with a single susceptibility allele. The value of 1·38 found here is characteristic of samples of IDDM individuals where an excess of DR3, 4's is often observed thus suggesting a more complex mode of inheritance for susceptibility (Falk, 1984). The 'control group' was made up of the parental haplotype pairs not present in the affected child (only families in which all four HLA haplotypes could be followed were used). There were 146 parental control haplotypes. The allele frequencies for *DR3*, *DR4*, and *DRX* in this group were 0·103, 0·082, and 0·815 respectively. These values agree remarkably well with the total frequencies obtained for the 'random mating population' comprising all caucasoid random individuals submitted to the 9th HLA Workshop (Baur *et al.* 1984) (see. e.g. the table on page 694, where the *DR* marginal frequencies are 0·122, 0·129, and 0·749 for the same three *DR* alleles). If the control haplotypes from each family are assumed to be a 'control individual', we obtain a control population sample of 73 which is in H–W equilibrium ($\chi^2 = 1{\cdot}79$, 2 d.f., see Table 2).

In Table 3, we compare the HRR's for DR3 and DR4 to the RR's calculated using a 'contrived control population' from the 9th HLA Workshop population data referred to above. This 'population' is assumed to be in H–W equilibrium and our 'random sample' is of the same

Table 3. *HRR's and RR's for the DR3 and DR4 antigens in a sample of simplex IDDM patients*

(The control for the HRR's is the sample of parental haplotypes not present in the affected individuals. The control for the RR's was obtained by 'creating' a H-W sample assuming the antigen frequencies recorded for the 9th HLA workshop (Baur *et al.* 1984).)

	HRR				RR		
	DR3				*DR3*		
	+	−			+	−	
Disease	47	43	90	Disease	47	43	90
control	15	58	73	control	21	69	90
	62	101	163		68	112	180
	HRR = 4·23				RR = 3·59		
	$p = 2{\cdot}6 \times 10^{-5}$				$p = 5{\cdot}3 \times 10^{-5}$		
	DR4				*DR4*		
	+	−			+	−	
Disease	58	32	90	Disease	58	32	90
control	12	61	73	control	22	68	90
	70	93	163		80	100	180
	HRR = 9·21				RR = 5·60		
	$p = 7{\cdot}6 \times 10^{-10}$				$p = 6{\cdot}8 \times 10^{-8}$		

Table 4. *HRR's and RR's for the DR3, 3, DR3, 4 and DR4, 4 phenotypes*

(Samples are the same as those described in Table 3. In each case comparison is made relative to the 'base group' DRX, X to avoid the problems of non-independent risk estimates.).

DR type	Disease sample	Parental control	Workshop control
DR3, 3	6	0	1·3
DR3, 4	25	2	2·8
DR4, 4	4	0	1·5
DR3, X	16	13	16·4
DR4, X	29	10	17·4
DRX, X	10	48	50·5
Total	90	73	89·9

HRR	RR
HRR(3, 4) = 60·0	RR(3, 4) = 45·1
HRR(3, 3) = undefined	RR(3, 3) = 23·3
HRR(4, 4) = undefined	RR(4, 4) = 13·5

If 'expected values' are substituted for the zero observations in the parental control, one gets:

HRR′(3, 3) = 37·4,
HRR′(4, 4) = 39·2.

size as our disease sample (i.e. 90 individuals). Table 4 gives HRR's and RR's for the three DR phenotypes DR3, 3, DR3, 4, and DR4, 4 using the same samples. Here the risks are compared to the baseline phenotype DRX, X in each case since the risks are not independent (cf. Curie-Cohen, 1981, Svejgaard & Ryder, 1981). Note that the HRR's for DR3, 3 and DR4, 4 are undefined since there are no 'individuals' with those phenotypes in the control sample of 73. If expected values are substituted for the 'zero' values in those cases HRR's can be estimated as given at the bottom of Table 4, but the use of such estimates must be made with caution.

DISCUSSION

One of the major problems inherent in proper calculations of relative risks (RR's) is that of choosing an appropriate control. A basic assumption in the use of RR's is that both the affected sample and the control sample are chosen at random from the same genetically homogeneous random mating population with no selection criteria except for the disease status required for inclusion in the affected sample. In practice this is a difficult criterion to fulfil. Additionally, it adds a significant amount of work to select and test such a control sample. It is therefore often assumed that the control sample is simply a hypothetical sample created from a population thought to be similar to that of the disease sample and 'generated' from that population by assuming H–W equilibrium and some reasonable sample size (cf. Svejgaard & Ryder, 1981, and our 'contrived' sample of the previous section).

Given the known heterogeneity of current urban populations, even within the less heterogeneous European countries, use of population control data culled, for example, from HLA workshop surveys, may alter the significance of calculated RR's. Although, in the examples given here the results are significant for both RR's and HRR's (Table 3), the 'p-values' for significance differ by two-fold (for *DR3*) and 100-fold (for *DR4*), with the HRR's being more significant in each case. If less extreme samples were tested, careless choice of the control group could very well make the difference between statistical significance and non-significance (resulting in either a type I or a type II error).

Methods have previously been proposed for using sibship information to calculate 'risks'. For example, Clarke (1961) describes a method, attributed to C. A. B. Smith, for using sibships to test for a significant risk of duodenal ulcers to individuals of blood group O. The method used is somewhat different from that described here in that an observed and expected probability of being group O is assigned to the propositus in each sibship where the expected value depends on the makeup of the sibship. The significance is then based on a comparison of pooled observed and expected values over a set of sibships. This method does overcome the problem of heterogeneity but, because of the way the test is constructed, only a small part of the data can be used. In Clarke's example, therefore, the associations found when using the general population as a control were very much decreased when using Smith's sibship method. This does not seem to be the case using HRR's where the associations remain strong.

By using the two parental haplotypes not present in the single diseased individuals of the disease sample as the control 'sample', we are assured of having both samples from the same genetic population and, as was demonstrated above, this sample should represent a random sample of haplotype pairs (or 'individuals') from that population. Care must still be taken to ensure that the population chosen is genetically homogeneous, to the extent possible, but the task of obtaining an appropriate control is simplified.

If the disease is dominant rather than recessive, the HRR can still be used in the same way. Although it is not known whether the disease allele is present on the paternal haplotype ('a') or the maternal ('c') or perhaps on both, the other two parental haplotypes, b and d, will still represent random haplotypes from the underlying population, provided that the conditions mentioned for the recessive case obtain.

If a family is selected through more than one affected child, the situation is somewhat different. If the two affected sibs share the same two HLA haplotypes then the other two should still represent random haplotypes from the population. However, if they share fewer than two haplotypes, the situation is more complicated. Now three (or possibly four) haplotypes are known to carry the disease allele in the recessive case. If the disease is dominant, it is possible, but not certain, that a single shared haplotype carries the disease allele. If no haplotype is shared, it is not possible to define disease-carrying haplotypes with certainty. In such cases it would therefore be difficult to define a control sample of random haplotypes meeting the necessary criteria.

Two other points should be emphasized. If there is differential selection between genotypes at the susceptibility locus, (e.g. reduced fertility) a bias might be introduced such that the control haplotypes could no longer be considered a random population sample. Thus we require compliance with assumption (3) of our model to ensure the proper distribution of susceptibility alleles in the 'control' haplotypes.

Further, if the population from which the sample is drawn is genetically heterogeneous with respect to the disease, the HRR as well as the RR may be difficult to interpret as well as to use. In an extreme case a population might be made up of two ethnically distinct subpopulations that do not interbreed. Assume that the disease of interest occurs in only one of two such subpopulations. An estimate of the HRR would come entirely from a sample taken from the subpopulation where the disease is present and would be relevant only to that population (individuals in the other group having no risk, by definition). On the other hand, the RR would assign a risk over the entire population that would be too low for individuals in the susceptible part of the population and too high for individuals in the non-susceptible part.

We wish to thank Drs Jurg Ott, Neil Risch and C. A. B. Smith for helpful and constructive comments on an earlier draft of this paper.

This work was supported by NIH grant GM29177.

REFERENCES

Baur, M. P., Neugebauer, M. & Albert, E. D. (1984). Reference tables of two-locus haplotype frequencies for all MHC marker loci. In *Histocompatibility Testing* (eds. E. D. Albert, M. P. Baur and W. R. Mayr), pp. 677–755. Berlin: Springer-Verlag.

Bertrams, J. & Baur, M. P. (1984). Insulin-dependent diabetes mellitus. In *Histocompatibility Testing* (eds. E. D. Albert, M. P. Baur and W. R. Mayr), pp. 348–358. Berlin: Springer-Verlag.

Clarke, C. A. (1961). Blood Groups and Disease. *Progress in Medical Genetics* **1**, 81–119.

Curie-Cohen, M. (1981). HLA antigens and susceptibility to juvenile diabetes: do additive relative risks imply genetic heterogeneity? *Tissue Antigens* **17**, 136–148.

Falk, C. T., Mendell, N. R. & Rubinstein, P. (1983). Effect of population associations and reduced penetrance on observed and expected genotype frequencies in a simple genetic model: application to HLA and insulin dependent diabetes mellitus. *Ann. Hum. Genet.* **47**, 161–165.

Falk, C. T. (1984). A two-susceptibility-allele model for genetic diseases and associated marker loci: differences and similarities to a one-*s*-allele model. *Ann. Hum. Genet.* **48**, 87–95.

Rubinstein, P., Walker, M., Carpenter, C., Carrier, C., Krassner, J., Falk, C. & Ginsberg, F. (1981). Genetics of HLA disease associations. The use of the haplotype relative risk (HRR) and the "haplo-delta" (Dh) estimates in juvenile diabetes from three racial groups. *Human Immunology* **3**, 384 (Abstract).

Svejgaard, A. & Ryder, L. P. (1981). HLA genotype distribution and genetic models of insulin-dependent diabetes mellitus. *Ann. Hum. Genet.* **45**, 293–298.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Ann. Hum. Genet.* **19**, 251–253.